PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04N — A2

(11) International Publication Number: **WO 96/31047**

(43) International Publication Date: 3 October 1996 (03.10.96)

(21) International Application Number: PCT/US96/04400

(22) International Filing Date: 29 March 1996 (29.03.96)

(30) Priority Data:

08/414,437	31 March 1995 (31.03.95)	US
08/554,848	7 November 1995 (07.11.95)	US

(71) Applicant *(for all designated States except US)*: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA [US/US]; 22nd floor, 300 Lakeside Drive, Oakland, CA 94612-3550 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: JAIN, Ramesh [US/US]; 4715 Reedly Terrace, La Jolla, CA 92130 (US). WAKIMOTO, Koji [JP/US]; 6-36-8-204, Shonandai, Fujisawa, Kanagawa 252 (JP). MOEZZI, Saied [US/US]; 10420 Caminito Alvarez, San Diego, CA 92126 (US). KATKERE, Arun [IN/US]; 9500 Gilman Drive, La Jolla, CA 92093-0407 (US).

(74) Agent: FUESS, William, C.; Suite II-G, 10951 Sorrento Valley Road, San Diego, CA 92121-1613 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, US, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

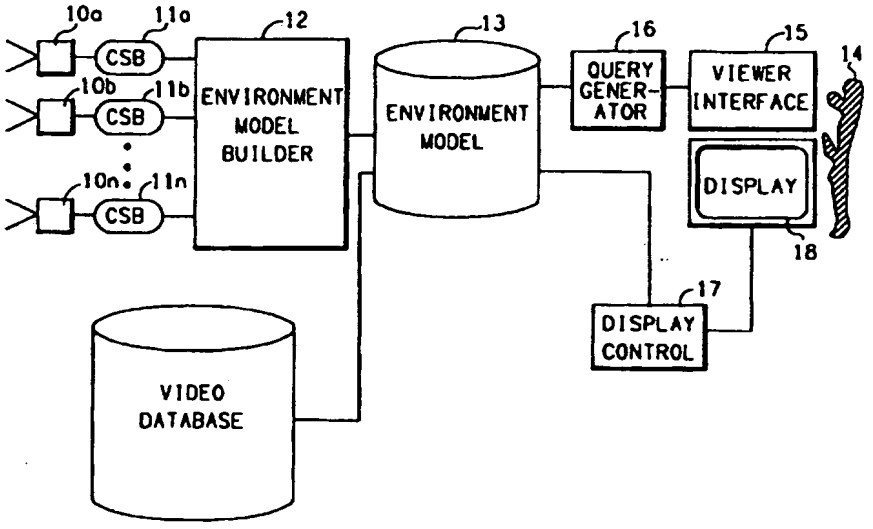
Published

Without international search report and to be republished upon receipt of that report.

(54) Title: IMMERSIVE VIDEO

(57) Abstract

Immersive video, or television, images of a real-world scene are synthesized (i) on demand, (ii) in real time, (iii) as linked to any of a particular perspective on the scene, or an object or event in the scene, (iv) in accordance with user-specified parameters of presentation, including panoramic or magnified presentations, and/or (v) stereoscopically. The synthesis of virtual images is based on computerized video processing -- called "hypermosaicing" -- of multiple live video perspectives on the scene. In hypermosaicing a knowledge database contains information about the scene; for example scene geometry, shapes and behaviors of objects in the scene, and/or internal and/or external camera calibration models. Multiple video cameras each at a different spatial location produce multiple two-dimensional video images of the scene. A viewer/user specifies viewing criterion (ia) at a viewer interface. A computer, typically one or more engineering work station class computers or better, includes in software and/or hardware (i) a video data analyzer for detecting and for tracking scene objects and their locations, (ii) an environmental model builder combining multiple scene images to build a 3-D dynamic model recording scene objects and their instant spatial locations, (iii) a viewer criterion interpreter, and (iv) a visualizer for generating from the 3-D model in accordance with the viewing criterion one or more particular 2-D video image(s) of the scene. A video display receives and displays the synthesized 2-D video image(s). Nonetheless to being built and maintained by use of simplifying assumptions, the 3-D dynamic model is powerful, flexible and useful in permitting diverse scene views.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgystan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic			SE	Sweden
CG	Congo	KR	Republic of Korea	SG	Singapore
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LR	Liberia	SZ	Swaziland
CS	Czechoslovakia	LT	Lithuania	TD	Chad
CZ	Czech Republic	LU	Luxembourg	TG	Togo
DE	Germany	LV	Latvia	TJ	Tajikistan
DK	Denmark	MC	Monaco	TT	Trinidad and Tobago
EE	Estonia	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	UG	Uganda
FI	Finland	ML	Mali	US	United States of America
FR	France	MN	Mongolia	UZ	Uzbekistan
GA	Gabon	MR	Mauritania	VN	Viet Nam

IMMERSIVE VIDEO

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention generally concerns (i) multimedia, (ii) video, including video-on-demand and interactive video, and (iii) television, including television-on-demand and interactive television.

The present invention particularly concerns automated dynamic selection of one video camera/image from multiple real video cameras/images in accordance with a particular perspective, an object in the scene, or an event in the video scene.

The present invention also particularly concerns the synthesis of diverse spatially and temporally coherent and consistent virtual video cameras, and virtual video images, from multiple real world video images that are obtained by multiple real video cameras.

The present invention still further concerns the creation of three-dimensional video image databases, and the location and dynamical tracking of video images of selected objects depicted in the databases for, among other purposes, the selection of a real camera or image, or the synthesis of a virtual camera or image, best showing the selected object.

The present invention still further concerns (i) interactive synthesis of video, or television, images of a real-world scene on demand, (ii) the synthesis of virtual video images of a real-world scene in real time, or virtual television, (iii) the synthesis of virtual video images/virtual television pictures of a real-world scene which video images/virtual television are linked to any of a particular perspective on the video/television scene, an object in the video/television scene, or an event in the video/television scene, (iv) the synthesis of virtual video images/virtual television pictures of a real-world scene wherein the pictures are so synthesized to user-specified parameters of presentation, e.g. panoramic, or at magnified scale if so desired by the user, and (v) the synthesis of 3-D stereoscopic virtual video images/virtual television.

2. Description of the Prior Art

2.1 Limitations in the Present Viewing of Video and Television

The traditional model of television and video is based on a single video stream transmitted to a passive viewer. A viewer has the option to watch the particular video stream, and to re-

watch should the video be recorded, but little else. Due to the emergence of the information highways and other related information infrastructure circa 1995, there has been considerable interest in concepts like video-on-demand, interactive movies, interactive TV, and virtual presence. Some of these concepts are exciting, and suggest many dramatic changes in society due to the continuing dawning of the information age.

It will shortly be seen that this specification teaches that a novel form of video, and television, is possible where a viewer of video, or television, depicting a real-world scene may select a particular perspective from which perspective the scene will henceforth be presented. The viewer may alternatively select a particular object -- which may be a dynamically moving object -- or even an event in the real world scene that is of particular interest. As the scene develops then it will be presented to the viewer with the selected object or the selected event (if occurring) prominently featured.

Accordingly, video presentation of a real-world scene in accordance with the present invention will be seen to be interactive with both (i) a viewer of the scene and, in the case of a selected dynamically moving object, or an event, in the scene, (ii) the scene itself. True interactive video or television is thus presented to a viewer.

In an extension of the present invention the image presented to the viewer will be seen to be a virtual image that is not mandated to correspond to any real world camera nor to any real world camera image. A viewer may thus view a video or television of a real-world scene from a vantage point (i.e., a perspective on the video scene), and/or dynamically in response to objects moving in the scene and/or events transpiring in the scene, in manner that is not possible in reality. The viewer may, for example, view the scene from a point in the air above the scene, or from the vantage point of an object in the scene, where no real camera exists or even, in some cases, can exist.

This video system, and approach, is called Multiple Perspective Interactive ("MPI") video. MPI video will be seen to be the basis, and the core, of an even more sophisticated "immersive video" (non-real-time and "immersive telepresence" or "Visualized Reality (VisR) (real-time) system of the present invention.

MPI video supports the editing of, and viewer interaction with, video and television in a manner that is useful in viewing activities ranging from education to entertainment. In particular, in conventional video, viewers are substantially passive; all they can do is to control the flow of video by pressing buttons such as play, pause, fast forward or fast

reverse. These controls essentially provide the viewer only one choice for a particular segment of video: the viewer can either see the video (albeit at a controllable rate), or skip it.

In the case of live television broadcast, viewers have essentially no control at all. A viewer must either see exactly what a broadcaster chooses to show, or else change away from that broadcaster and station. Even in sports and other broadcast events where multiple cameras are used, a viewer has no choice except the obvious one of either viewing the image presented or else using a remote control so as to "surf" multiple channels.

Interactive video and television systems such as MPI video make good use of the availability of increased video bandwidth due to new satellite and fiber optic video links, and due to advances in several areas of video technology. Author George Gilder argues that because the viewers really have no choice in the current form of television, it is destined to be replaced by a more viewer-driven system or device. See George Gilder; *Life After Television: The coming transformation of Media and American Life*, W. W. Norton & Co., 1994.

The related invention of MPI video makes considerable progress -- even by use of currently existing technology -- towards "liberating" video and TV from the traditional single-source, broadcast, model, and towards placing each viewer in his or her own "director's seat".

A three-dimensional (3-D) video model, or database, is used in MPI video. The immersive video and immersive telepresence systems of the present invention preserve, expand, and build upon this 3-D model. This three-dimensional model, and the functions that it performs, are well and completely understood, and will be completely taught within this specification. However, the considerable computational power required if a full custom virtual video image for each viewer is to be synthesized in real time and on demand requires that the model should be constructed and maintained in consideration of (1) powerful organizing principles, (ii) efficient algorithms, and (iii) effective and judicious simplifying assumptions. This then, and more, is what the present invention will be seen to concern.

2.2 Previous Scene-Interactive Video and Television

Existing scene-interactive video and television is nothing so grandiose as permitting a user/viewer to interact with the objects and/or events of a scene -- as will be seen to be the subject of the present and related inventions. Rather, the interaction with the scene is simply that of a machine -- a computer -- that must recognize, classify and, normally, adapt its responses to what it "sees" in the scene. Scene-interactive

video and television is thus simply an extension of machine vision so as to permit a computer to make decisions, sound alarms, etc., based on what it detects in, and detects to be transpiring in, a video scene. Two classic problems in this area (which problems are not commensurate in difficulty) are (i) security cameras, which must detect contraband, and (ii) an autonomous computer-guided automated battlefield tank, which must sense and respond to its environment.

The general concepts, and voluminous prior art, concerning "machine vision", "(target) classification", and "(target) tracking" are all relevant to the present invention. However, the video and television systems of the present invention -- while doing very, very well in each of viewing, classifying and tracking, will be seen to come to these problems from a very different perspective than does the prior art. Namely, the prior art considers platforms -- whether they are rovers or warships -- that are "located in the world", and that must make sense of their view thereof from essentially but a single perspective centered on present location.

The present invention functions oppositely. It "defines the world", or at least so much of the world is "on stage" and in view to (each of) multiple video cameras. The video and television systems of the present invention have at their command a plethora of correlatable and correlated, simultaneous, positional information. Once it is known where each of multiple cameras are, and are pointing, it is a straightforward matter for computer processes to fix, and to track, items in the scene.

The systems, including the MPI-video subsystem, of the present invention will be seen to perform co-ordinate transformation of (video) image data (i.e., pixels), and to do this during a generation of two- and three-dimensional image databases.

2.3 Previous Composite Video and Television

The present invention of immersive video will be seen to involve the manipulation, processing and compositing of video data in order to synthesize video images. (Video compositing is the amalgamation of video data from separate video streams.) It is known to produce video images that -- by virtue of view angle, size, magnification, etc. -- are generally without exact correspondence to any single "real-world" video image. The previous process of so doing is called "video mosaicing".

The underlaying task in video mosaicing is to create larger images from frames obtained from one or more single cameras, typically one single camera producing a panning video stream. To generate seamless video mosaics, registration and alignment of the frames from a sequence are critical issues.

Next, the immersive video system of the present invention will be seen to use its several streams of 2D video data to build and maintain a 3D video database. The utility of such 3D database in the synthesis of virtual video images seems clear. For example, an arbitrary planar view of the scene will contain the data of 2D planar slice "through" the 3D database.

The limitation on such a scheme of a information-intensive representation, and manipulation, of the video data of a real-world scene is that a purely "brute force" approach is impossible with presently available technology. The "trade-off" in handling a lot of video data is that (i) certain scene (or at least scene video) constraints must be imposed, (ii) certain simplifying assumptions must be made (regarding the content of the video information, (iii) certain expediencies must be embraced (regarding the manipulations of the video data), and/or (iv) certain limitations must be put on what images can, or cannot, be synthesized from such data. (The present invention will be seen to involve essentially no (iv) limitations on presentation.) Insofar as the necessary choices and trade-offs are astutely made, then it may well be possible to synthesize useful and aesthetically pleasing video, and even television, images by the use of tractable numbers of affordable computers and other equipments running software programs of reasonable size.

The immersive video system of the present invention will so show that -- (i) certain scene constraints being made, (ii) certain simplifying assumptions being made regarding scene objects and object dynamical motions being made, and (iii) certain computational efficiencies in the manipulations of video data being embraced -- it is indeed possible, and even practical, to so synthesize useful and aesthetically pleasing video, and even television, images.

SUMMARY OF THE INVENTION

1. Machine Dynamic Selection, of One Video Camera/Image of a Scene from Multiple Video Cameras/Images of the Scene in Accordance with a Particular Perspective on the Scene, an Object in the Scene, or an Event in the Scene

The present invention contemplates machine dynamic selection, of one video camera/image of a scene from multiple video cameras/images of the scene in accordance with a particular perspective on the scene, an object in the scene, or an event in the scene.

The present invention thus contemplates making each and any viewer of a video or a television scene to be his or her own

proactive editor of the scene, having the ability to interactively dictate and select -- in advance of the unfolding of the scene, and by high-level command -- a particular perspective by which the scene will be depicted, as and when the scene unfolds.

The viewer can command the selection of real, or -- in advanced embodiments of the invention -- even the synthesis of virtual, video images of the scene in response to any of his or her desired and selected (i) spatial perspective on the scene, (ii) static or dynamically moving object appearing in the scene, or (iii) event depicted in the scene. The viewer -- any viewer -- is accordingly considerably more powerful than even the broadcast video editor of, for example, a live sporting event circa 1995. The viewer is accorded the ability to (i) select in advance a preferred video perspective of view as optionally may be related to dynamic object movements and/or to events unfolding in the scene, and even, as the ultimate extension of the invention, (ii) to synthesize video views where no real video camera even exists.

1.1 The Basis of the Present Invention in Multiple Perspective Interactive (MPI) Video

The basis, and most basic part, of the present invention is called Multiple Perspective Interactive, or MPI, Video. MPI Video forms the core of the Immersive Video discussed hereinafter in section 3.

For example, in accordance with the present invention of MPI Video a viewer of an American football game on video or on television can command a consistent "best" view of (i) one particular player, or, alternatively (ii) the football itself as will be, from time to time, handled by many players. The system receives and processes multiple video views (images) generally of the football field, the football and the players within the game. The system classifies, tags and tracks objects in the scene, including static objects such as field markers, and dynamically moving objects such as the football and the football players. Some of the various views (images) will at times, and from time to time, be "better" -- by various criteria -- in showing certain things than are other views.

In the rudimentary embodiment of the invention taught within this specification the system will consistently, dynamically, select and present a single "best" view of the selected object (for example, the football, or a particular player). This will require, and the system will automatically accomplish, a "handing off" from one camera to another camera as different ones of multiple cameras best serve to image over time the selected object. In the ultimate extension of the present

invention, the viewer can ask to be shown a synthesized video view, such as from a perspective constantly positioned behind a certain offensive running back, where no real video camera actually exists.

The system of the invention is powerful (i) in accepting viewer specification at a high level of those particular objects and/or events in the scene that the user/viewer desires to be shown, and (ii) to subsequently identify and track all user/viewer-selected objects and events (and still others for other users/viewers) in the scene.

The system of the present invention can also, based on its scene knowledge database, serve to answer questions about the scene.

Finally, the system of the present invention can replay events in the scene from the same perspective, or from selected new perspectives, depending upon the desires of the user/viewer. It is not necessary for the user/viewer to "find" the best and proper image; the system performs this function. For example, if the user/viewer wants to see how player number twenty (#20) came to make an interception in the football game, then he or she could order a replay of the entire down focused on player number twenty (#20).

For example, and continuing with the example of an American football game, an individual viewer can ask questions like: Who is the particular player shown marked by my cursor? Where is player Mr. X? Where is the football?

In advanced, image-synthesizing, embodiments of the system of the present invention, the user/viewer can generate commands like: "replay for me at 1/2 speed the event of the fumble as shown from a straight overhead view". Such commands are honored by the system of the present invention even though no real video camera may, in actuality, exist at this precise overhead location.

1.2 Machine Dynamic Selection, of One Video Camera/Image of a Scene from Multiple Video Cameras/Images of the Scene in Accordance with a Particular Perspective on the Scene, an Object in the Scene, or an Event in the Scene

The present invention contemplates selecting real, or -- in advanced embodiments -- synthesizing virtual, video/television images of a scene from multiple real video/television images of the scene, particularly so as to select or to synthesize video/television images that are linked to any such (i) spatial perspective(s) on the scene, (ii) object(s) in the scene, or (iii) event(s) in the scene, as are selectively desired by a user/viewer to be shown.

The method of the invention is directed to presenting to a

user/viewer a particular, viewer-selected, two-dimensional video image of a real-world, three-dimensional, scene. In order to do so, multiple video cameras, each at a different spatial location, produce multiple two-dimensional images of the real-world scene, each at a different spatial perspective. Objects of interest in the scene are identified and classified in these two-dimensional images. These multiple two-dimensional images of the scene, and their accompanying object information, are then combined in a computer into a three-dimensional video database, or model, of the scene. The database is called a model because it incorporates information about the scene as well as the scene video. It incorporates, for example, a definition, or "world view", of the three-dimensional space of the scene. The model of a football game knows, for example, that the game is played upon a football field replete with static, fixed-position, field yard lines and hash mark markings, as well as of the existence of the dynamic objects of play. The model is, it will be seen, not too hard to construct so long as there are, or are made to be, sufficient points of reference in the imaged scene. It is, conversely, almost impossible to construct the 3-D model, and select or synthesize the chosen image, of an amorphous scene, such as the depths of the open ocean. (Luckily, viewers are generally more interested in people in the world than in fish.) The computer also receives from a prospective user/viewer of the scene a user/viewer-specified criterion relative to which criterion the user/viewer wishes to view the scene.

From the (i) 3-D model and (ii) the criterion, the computer produces a particular two-dimensional image of the scene that is in accordance with the user/viewer-specified criterion. This particular two-dimensional image of the real-world scene is then displayed on a video display to the user/viewer.

At the highest-level, the description of the previous paragraphs regarding the method of the present invention, and the computer-based system performing the method, may not seem much different in effect than that prior art system presently accorded, say, a network sports director who is able to select among many video feeds in accordance with his (or her) own "user/viewer-specified criterion". The significance of the production of the three-dimensional video model (of the real-world scene) by the method, and in the system, of the present invention is, at this highest level of describing the system's functions, as yet unclear. Consider, then, exactly what flows from the method, and the system, of the present invention that produces and uses a three-dimensional video model.

First, the computer may ultimately produce, and the display may finally show, only such a particular two-dimensional image

of the scene -- in accordance with the user/viewer-specified criterion -- as was originally one of the images of the real-world scene that was directly imaged by one of the multiple video cameras. This is, indeed, the way the rudimentary embodiment of the invention taught and shown herein functions. At first consideration, this automatic camera selection may seem unimpressive. However, consider not only that the user/viewer criterion is specified at a high level, but that the appropriate, selected, scene image may change over time in accordance with just what is imaged, and in what location(s), by which camera(s), and in accordance with just what transpires in the scene. In other words, the evolving contents of the scene, as the scene is imaged by the multiple cameras and as it is automatically interpreted by the computer, determine just what image of the scene is shown at any one time, and just what sequence of images are shown from time to time, to the user/viewer. Action in the scene "feeds back" on how the scene is shown to the viewer!

Second, in advanced embodiments of the system, the computer is not limited to selecting from the three-dimensional model a two-dimensional image that is, or that corresponds to, any of the images of the real-world scene as was imaged by any of the multiple video cameras. Instead, the computer may synthesize from the three-dimensional model a completely new two-dimensional image that is without exact equivalence to any of the images of the real-world scene as have been imaged by any of the multiple video cameras.

Third, the user/viewer-specified criterion may be of a particular spatial perspective relative to which the user/viewer wishes to view the scene. This spatial perspective need not be immutably fixed, but can instead be linked to a dynamic object in the scene. In the case of generating a scene view from a user/viewer-specified spatial perspective, the computer produces from the three-dimensional model a particular two-dimensional image of the scene that is in best accordance with some particular spatial perspective criterion that has been received from the viewer. The particular two-dimensional image of the scene that is generated and displayed may, or may not, be, or be equivalent to, any real image of the scene as was obtained by any of the video cameras. In other words, in advanced embodiments of the invention the scene image shown may be a virtual image. Even if the image shown is a real image, the computer will still automatically select, and the display will still display, over time, those actual images of the scene as are imaged, over time, by different ones of the multiple video cameras. Automated scene switching, especially in relation to dynamic objects in the scene, is not known to the inventors to

exist in the prior art.

Fourth, the user/viewer-specified criterion may be of a particular object in the scene. In this case the computer will combine the images from the multiple video cameras not only so as to generate a three-dimensional video model of the scene, but so as to generate a model in which objects in the scene are identified. The computer will subsequently produce, and the display will subsequently show, the particular image -- whether real or virtual -- appropriate to best show the selected object. Clearly this is a feedback loop: the location of an object in the scene serves to influence, in accordance with a user/viewer selection of the object, how the scene is shown. Clearly the same video scene could be, if desired, shown over and over, each time focusing view on a different selected object in the scene.

Moreover, the selected object may either be static, and unmoving, or dynamic, and moving, in the scene. Regardless of whether the object in the scene is static or dynamic, it is preferably specified to the system by the user/viewer by act of positioning a cursor on the video display. The cursor is a special type that unambiguously specifies an object in the scene by an association between the object position and the cursor position in three dimensions, and is thus called "a three-dimensional cursor".

Fifth, the criterion specified by the user/viewer may be of a particular event in the scene. In this case the computer will again combine the images from the multiple video cameras not only so as to generate a three-dimensional video model of the scene, but so as to generate a model in which one or more dynamically occurring event(s) in the scene are recognized and identified. The computer will subsequently produce, and the display will show, a particular image -- whether real or virtual -- that is appropriate to best show the selected event. Clearly this is again a feedback loop: the location of an event in the scene influences, in accordance with a viewer selection of the event, how the scene is shown.

Sixth, and finally, the method of the invention may be performed in real time as interactive television. The television scene will be presented to a user/viewer interactively in accordance with the user/viewer-specified criterion.

2. Immersive Video, Also Called Telepresence, Also Called Visual Reality (VisR)

The present invention still further contemplates telepresence and immersive video, being the non-real-time creation of a synthesized, virtual, camera/video image of a real-world scene, typically in accordance with one or more

viewing criteria that are chosen by a viewer of the scene. Immersive video, or telepresence, or visual reality (VisR) is an extension of Multiple Perspective Interactive (MPI) video.

In immersive video the creation of the virtual image is based on a computerized video processing -- in a process called hypermosaicing -- of multiple video views of the scene, each from a different spatial perspective on the scene.

When the synthesis and the presentation of the virtual image transpires as the viewer desires -- and particularly as the viewer indicates his or her viewing desires simply by action of moving and/or orienting any of his or her body, head and eyes -- then the process is called "immersive telepresence", or simply "telepresence". Alternatively, the process is sometimes called "visual reality", or simply "VisR".

(The proliferation of descriptive terms has more to do with the apparent reality(ies) of the synthesized views drawn from the real-world scene than it does with the system and processes of the present invention for synthesizing such views. For example, a quite reasonable ground level view of a football quarterback as is may be synthesized by the system and method of the present invention may appear to a viewer to have been derived from a hand-held television camera, although in fact no such camera exists and the view was not so derived. These views of common experience are preliminarily called "telepresence". Contrast a magnified, eye-to-eye, view with an ant. This magnified view is also of the real-world, although it is clearly a view that is neither directly visible to the naked eye, nor of common experience. Although derived by entirely the same processes, views of this latter type of synthesized view of the real world is preliminarily called "visual reality", or "VisR", by juxtaposition of such views the similar sensory effects engendered by "virtual reality", or "VR".)

2.1 Telepresence, Both Immersive and Interactive

In one of its aspects, the present invention is embodied in a method of telepresence, being a video representation of being at real-world scene that is other than the instant scene of the viewer. The method includes (i) capturing video of a real-world scene from each of a multiplicity of different spatial perspectives on the scene, (ii) creating from the captured video a full three-dimensional model of the scene, and (iii) producing, or synthesizing, from the three-dimensional model a video representation on the scene that is in accordance with the desired perspective on the scene of a viewer of the scene.

This method is thus called "immersive telepresence" because the viewer can view the scene as if immersed therein, and as if present at the scene, all in accordance with his or her desires.

Namely, it appears to the viewer that, since the scene is presented as the viewer desires, the viewer is immersed in the scene. Notably, the viewer-desired perspective on the scene, and the video representation synthesized in accordance with this viewer-desired perspective, need not be in accordance with any of the video captured from any scene perspective.

The video representation can be in accordance with the position and direction of the viewer's eyes and head, and can exhibit "motional parallax". "Motional parallax" is normally and conventionally defined as a three-dimensional effect where different views on the scene are produced as the viewer moves position, making the viewer's brain to comprehend that the viewed scene is three-dimensional. Motional parallax is observable even if the viewer has but one eye.

Still further, and additionally, the video representation can be stereoscopic. "Stereoscopy" is normally and conventionally defined as a three-dimensional effect where each of the viewer's two eyes sees a slightly different view on the scene, thus making the viewer's brain to comprehend that the viewed scene is three-dimensional. Stereoscopy is detectable even should the viewer not move his or her head or eyes in spatial position, as is required for motional parallax.

In another of its aspects, the present invention is embodied in a method of telepresence where, again, video of a real-world scene is obtained from a multiplicity of different spatial perspectives on the scene. Again, a full three-dimensional model of the scene is created the from the captured video. From this three-dimensional model a video representation on the scene that is in accordance with a predetermined criterion -- selected from among criteria including a perspective on the scene, an object in the scene and an event in the scene -- is produced, or synthesized.

This embodiment of the invention is thus called "interactive telepresence" because the presentation to the viewer is interactive in accordance with the criterion. Again, the synthesized video presentation of the scene in accordance with the criterion need not be, and normally is not, equivalent to any of the video captured from any scene perspective.

In this method of viewer-interactive telepresence the video representation can be in accordance with a criterion selected by the viewer, thus viewer-interactive telepresence. Furthermore, the presentation can be in accordance with the position and direction of the viewer's eyes and head, and will thus exhibit motional parallax; and/or the presentation can exhibit stereoscopy.

2.2 A System for Generating Immersive Video

A huge range of heretofore unobtainable, and quite remarkable, video views may be synthesized in accordance with the present invention. Nonetheless that an early consideration of exemplary video views of diverse types would likely provide significant motivation to understanding the construction, and the operation, of the immersive video system described in this section 2.2, discussion of these views is delayed until the next section 2.3. This is so that the reader, having gained some appreciation and understanding in this section 2.2 of the immersive video system, and process, by which the video views are synthesized, may later better place these diverse views in context.

An immersive video, or telepresence, system serves to synthesize and to present diverse video images of a real-world scene in accordance with a predetermined criterion or criteria. The criterion or criteria of presentation is (are) normally specified by, and may be changed at times and from time to time by, a viewer/user of the system. Because the criterion (criteria) is (are) changeable, the system is viewer/user-interactive, presenting (primarily) those particular video images (of a real-world scene) that the viewer/user desires to see.

The immersive video system includes a knowledge database containing information about the scene. Existence of this "knowledge database" immediately means that the something about the scene is both (i) fixed and (ii) known; for example that the scene is of "a football stadium", or of "a stage", or even, despite the considerable randomness of waves, of "a surface of an ocean that lies generally in a level plane". For many reasons -- including the reason that a knowledge database is required -- the antithesis of a real-world scene upon which the immersive video system of the present invention may successfully operate is a scene of windswept foliage in a deep jungle.

The knowledge database may contain, for example, data regarding any of (i) the geometry of the real-world scene, (ii) potential shapes of objects in the real-world scene, (iii) dynamic behaviors of objects in the real-world scene, (iv) an internal camera calibration model, and/or (v) an external camera calibration model. For example, the knowledge base of an American football game would be something to the effect that (i) the game is played essentially in a thick plane lying flat upon the surface of the earth, this plane being marked with both (yard) lines and hash marks; (ii) humans appear in the scene, substantially at ground level; (iii) a football moves in the thick plane both in association with (e.g., running plays) and detached from (e.g., passing and kicking plays) the humans; and (iv) the locations of each of several video cameras on the

football game are a priori known, or are determined by geometrical analysis of the video view received from each.

The system further includes multiple video cameras each at a different spatial location. Each of these multiple video cameras serves to produce a two-dimensional video image of the real-world scene at a different spatial perspective. Each of these multiple cameras can typically change the direction from which it observes the scene, and can typically pan and zoom, but, at least in the more rudimentary versions of the immersive video system, remains fixed in location. A classic example of multiple stationary video cameras on a real-world scene are the cameras at a sporting event, for example at an American football game.

The system also includes a viewer/user interface. A prospective viewer/user of the scene uses this interface to specify a criterion, or several criteria, relative to which he or she wishes to view the scene. This viewer/user interface may commonly be anything from head gear mounted to a boom to a computer joy stick to a simple keyboard. In ultimate applications of the immersive video system of the present invention, the viewer/user who establishes (and re-establishes) the criterion (criteria) by which an image on the scene is synthesized is the final consumer of the video images so synthesized and presented by the system. However, for more rudimentary present versions of the immersive video system, the control input(s) arising at the viewer/user interface typically arise from a human video sports director (in the case of an athletic event), from a human stage director (in the case of a stage play), or even from a computer (performing the function of a sports director or stage director). In other words, the viewing desires of the ultimate viewer/user may sometimes be translated to the immersive video system through an intermediary agent that may be either animate or inanimate.

The immersive video system includes a computer running a software program. This computer receives the multiple two-dimensional video images of the scene from the multiple video cameras, and also the viewer-specified criterion (criteria) from the viewer interface. At the present time, circa 1995, the typical computer functioning in an immersive video system is fairly powerful. It is typically an engineering work station class computer, or several such computers that are linked together if video must be processed in real time -- i.e., as television. Especially if the immersive video is real time -- i.e., as television -- then some or all of the computers normally incorporate hardware graphics accelerators, a well-known but expensive part for this class of computer. Accordingly, the computer(s) and other hardware elements of an

immersive video system are both general purpose and conventional but are, at the present time (circa 1995) typically "state-of-the-art", and of considerable cost ranging to tens, and even hundreds, of thousands of American dollars.

The system computer includes (in software and/or in hardware) (i) a video data analyzer for detecting and for tracking objects of potential interest and their locations in the scene, (ii) an environmental model builder for combining multiple individual video images of the scene to build a three-dimensional dynamic model of the environment of the scene within which three-dimensional dynamic environmental model potential objects of interest in the scene are recorded along with their instant spatial locations, (iii) a viewer criterion interpreter for correlating the viewer-specified criterion with the objects of interest in the scene, and with the spatial locations of these objects, as recorded in the dynamic environmental model in order to produce parameters of perspective on the scene, and (iv) a visualizer for generating, from the three-dimensional dynamic environmental model in accordance with the parameters of perspective, a particular two-dimensional video image of the scene.

The computer function (i) -- the video data analyzer -- is a machine vision function. The function can presently be performed quite well and quickly, especially if (i) specialized video digitalizing hardware is used, and/or (ii) simplifying assumptions about the scene objects are made. Primarily because of the scene model builder next discussed, abundant simplifying assumptions are both well and easily made in the immersive video system of the present invention. For example, it is assumed that, in a video scene of an American football game, the players remain essentially in and upon the thick plane of the football field, and do not "fly" into the airspace above the field.

The views provided by an immersive video system in accordance with the present invention not yet having been discussed, it is somewhat premature to explain how a scene object that is not in accordance with the model may suffer degradation in presentation. More particularly, the scene model is not overly particular as to what appears within the scene, but it is particular as to where within (the volume of) the scene an object to be modeled appears. Consider, for example, that the immersive video system can fully handle a scene-intrusive object that is not in accordance with prior simplifications -- for example, a spectator or many spectators or a dog or even an elephant walking onto a football field during or after a football game -- and can process these unexpected objects, and object movements quite as well as any other. However, if is necessary that the modeled object should

appear within a volume of the real-world scene whereat the scene model is operational -- basically that volume portion of the scene where the field of view of multiple cameras overlap. For example, a parachutist parachuting into a football stadium may not be "well-modeled" by the system when he/she is high above the field, and outside the thick plane, but will be modeled quite well when finally near, or on, ground level. By modeling "quite well", it is meant that, while the immersive video system will readily permit a viewer to examine, for example, the dentation of the quarterback if he or she is interested in staring the quarterback "in the teeth", it is very difficult for the system (especially initially, and in real time as television), to process through a discordant scene occurrence, such as the stadium parachutist, so well so as to permit the examination of his or her teeth also when the parachutist is still many meters above the field.

The computer function (ii) -- the environmental model builder -- is likely the "backbone" of the present invention. It incorporates important assumptions that, while scene specific, are generally of a common nature throughout all scenes that are of interest for viewing with the present invention.

In the first place, the environmental model is (i) three-dimensional (3-D), with both (i) static and (ii) dynamic components. The scene environmental model is not the scene image, nor the scene images rendered three-dimensionally. The current scene image, such as of the play action on a football field, may be, and typically is, considerably smaller than the scene environmental model which may be, for example, the entire football stadium and the objects and actors expected to be present therein. Within this three-dimensional dynamic environmental model both (i) the scene and (ii) all potential objects of interest in the scene are dynamically recorded as associated with, or "in", their proper instant spatial locations. (It should be remembered that the computer memory in which this 3-D model is recorded as actually one-dimensional (1-D), being but memory locations each of which is addressed by but a single 1-D address.) Understanding that the scene environmental model, and the representation of scene video information, in the present invention is 3-D will much simplify understanding of how the remarkable views discussed in the next section are derived.

At present there is not enough computer "horsepower" to process a completely amorphous unstructured video scene -- the windy jungle -- into 3-D, at least in real time (i.e., as television). It is, however, eminently possible to process many scenes of great practical interest and importance into 3-D if and when appropriate simplifying assumptions are made. In

accordance with the present invention, these necessary simplifying assumptions are very effective, making that production of the three-dimensional video database (in accordance with the 3-D environmental model) is very efficient.

First, the static "underlayment" or "background" of any scene is pre-processed into the three-dimensional video database. For example, the video model of an (empty) sports stadium -- the field, filed markings, goal posts, stands, etc. -- is pre-processed (as the environmental model) into the three-dimensional video database. From this point on only the dynamic elements in the scene -- i.e., the players, the officials, the football and the like -- need be, and are, dealt with. The typically greater portion of any scene that is (at any one time) static is neither processed nor re-processed from moment to moment, and from frame to frame. It need not be so processed or re-processed because nothing has changed, nor is changing. (In some embodiments of the immersive video system, the static background is not inflexible, and may be a "rolling" static background based on the past history of elements within the video scene.)

Meanwhile, dynamical objects in the scene -- which objects typically appear only in a minority of the scene (e.g. the football players) but may appear in the entire scene (e.g., the crowd) -- are preferably processed in two ways. If the computer recognition and classification algorithm can recognize -- in consideration of a priori model knowledge of items appearing in the scene such as the football, and the football players -- an item in the scene, then that item will be isolated, and will be processed/re-processed into the three-dimensional video database as a multiple voxel representation. (A voxel is a three-dimensional pixel.) Other dynamic elements of the scene that cannot be classified or isolated into the three-dimensional environmental model are swept up into the three-dimensional video database mostly in their raw, two-dimensional, video data form. Such a dynamic, but un-isolated, video element could be, for example, the movement of a crowd doing a "wave" motion at a sports stadium, or the surface of the sea.

As will be seen, those recognized and classified objects in the three-dimensional video database -- such as, for example, a football or a football player -- can later be viewed (to the limits of being obscured in all two-dimensional video data streams from which the three-dimensional video scene is composed) from any desired perspective. But it is not possible to view those unclassified and un-isolated dynamic elements of the scene that are stored in the 3-D video database in their 2-D video data from any random perspective. These dynamic objects can indeed be dynamically viewed, but it is impossible in the

system to, for example, go "behind" the moving crowd, or "under" the undulating surface of the sea.

The system and method does not truly know, of course, whether it is inserting into the instant three-dimensional video database in accordance with the scene environmental model an instant video image of a football quarterback taking a drink, or an instant video image of a football fan taking the same drink. Moreover, dynamic objects can both enter (e.g. as in coming onto the imaged field of play) and exit (e.g. as in leaving the imaged field of play) the scene. The system and method of the present invention for constructing a 3-D video scene deal only with (i) the scene environmental model, and (ii) the mathematics of the pixel dynamics. What must be recognized is that, in so doing, the system and method serve to discriminate between and among raw video image data in processing such image data into the three-dimensional video database.

These assumptions that the real-world scene contains both static and dynamic elements (indeed, preferably two kinds of dynamic elements), this organization, and these expediencies of video data processing are very important. They are collectively estimated to reduce the computational requirements for the maintenance of a 3-D video database of a typical real-world scene of genuine interest by a factor of from fifty to one hundred times (x 50 to x 100).

However, these simplifications have a price; thankfully normally one that is so small so as to be all but unnoticeable. Portions of the scene "where the action is, or has been" are entered into the three-dimensional video database quite splendidly. Viewers normally associate such "actions areas" with the center of their video or television presentation. When action spontaneously erupts at the periphery of a scene, it takes even our human brains -- whose attention has been focused elsewhere (i.e., at the scene center) -- several hundred milliseconds or so to recognize what has happened. So also, but in a different sense, it is possible to "sandbag" the system and method of the present invention by a spontaneous eruption of action, or dynamism, in a previously unclassified scene area. In a "first pass", or in real time (i.e., as television), the system and method of the present invention finds it hard to discriminate, and hard to process for entrance into the three-dimensional database, a three-dimensional scene object (or actor) where there was no previous scene object (or actor). Without a priori knowledge in the scene environmental model that a spectator may throw a bottle into a sporting arena, it is hard for the system of the present invention to classify and to process the throw and the thrower into the three-dimensional database so completely that the facial features of the thrower

may -- either upon an "instant replay" of the scene focusing on the area of the perpetrator or for that rare viewer who had been focusing his view to watch the crowd instead of the athletes all along -- immediately be recognized. (If the original raw video data streams still exist, then it is always possible to process them better.)

Finally, the algorithms themselves that are used to produce the three-dimensional video database are efficient.

Lastly, the system includes a video display that receives the particular two-dimensional video image of the scene from the computer, and that displays this particular two-dimensional video image of the real-world scene to the viewer/user as that particular view of the scene which is in satisfaction of the viewer/user-specified criterion (criteria).

2.3 Scene Views Obtainable With Immersive Video

To immediately note that a viewer/user of an immersive video system in accordance with the present invention may view the scene from any static or dynamic viewpoint -- regardless that a real camera/video does not exist at the chosen viewpoint -- only but starts to describe the experience of immersive video. Literally any video image(s) can be generated. The immersive video image(s) that is (are) actually displayed to the viewer/user are ultimately, in one sense, a function of the display devices, or the arrayed display devices -- i.e., the television(s) or monitor(s) -- that are available for the viewer/user to view. Because, at present (circa 1995), the most ubiquitous form of these display devices -- televisions and monitors -- have substantially rectangular screens, most of the following explanations of the various experiences of immersive video will be couched in terms of the planar presentations of these devices. However, when in the future new display devices such as volumetric three-dimensional televisions are built -- see, for example, U.S. Patent Nos. 5,268,862 and 5,325,324 each for a THREE-DIMENSIONAL OPTICAL MEMORY -- then the system of the present invention will stand ready to provide the information displayed by these devices.

2.3.1 Planar Video Views on a Scene

First, consider the generation of one-dimensional, planar and curved surface, video views on a scene.

Any "planar" view on the scene may be derived as the information which is present on any (straight or curved) plane (or other closed surface, such as a saddle) that is "cut" through the three-dimensional model of the scene. This "planar" surface may, or course, be positioned anywhere within the three-

dimensional volume of the scene model. Literally any interior or exterior virtual video view on the scene may be derived and displayed. Video views may be presented in any aspect ratio, and in any geometric form that is supported by the particular video display, or arrayed video displays (e.g., televisions, and video projectors), by which the video imagery is presented to the viewer/user.

Next, recall that a plane is but the surface of a sphere or cylinder of infinite radius. In accordance with the present invention, a cylindrical, hemispherical, or spherical panoramic view of a video scene may be generated from any point inside or outside the cylinder, hemisphere, or sphere. For example, successive views on the scene may appear as the scene os circumnavigated from a position outside the scene. An observer at the video horizon of the scene will look into the scene as if though a window, with the scene in plan view, or, if foreshortened, as if viewing the interior surface of a cylinder or a sphere from a peephole in the surface of the cylinder or sphere. In the example of an American football game, the viewer/user could view the game in progress as if he or she "walked" at ground level, or even as if he or she "flew at low altitude", around or across the field, or throughout the entire stadium.

A much more unusual panoramic cylindrical, or spherical "surround" view of the scene may be generated from a point inside the scene. The views presented greatly surpass the crude, but commonly experienced, example of "you are there" home video where the viewer sees a real-world scene unfold as a walking video cameraman shoots video of only a limited angular, and solid angular, perspective on the scene. Instead, the scene can be made to appear -- especially when the display presentation is made so as to surround the user as do the four walls of a room or as does the dome of a planetarium -- to completely encompass the viewer. In the example of an American football game, the viewer/user could view the game in progress as if he or she was a player "inside" the game, even to the extent of looking "outward" at the stadium spectators.

(It should be understood that where the immersive video system has no information -- normally because view is obscured to the several cameras -- than no image can be presented of such a scene portion, which portion normally shows black upon presentation. This is usually not objectionable; the viewer/user does not really expect to be able to see "under" the pile of football players, or from a camera view "within" the earth. Note, however, that when the 3-D video database does contain more than just surface imagery such as, for example, the complete 3-D human physiology (the "visible man"), then

"navigation" "inside" solid objects, into areas that have never been "seen" by eye or by camera, and at non-normal scales of view is totally permissible.)

Notably, previous forms of displaying multi-perspective, and/or surround, video presently (circa 1995) suffer from distortion. Insofar as the view caught at the focal plane of the camera, or each camera (whether film or video) is not identical to the view recreated for the viewer, the (often composite) views suffer from distortion, and to that extent a composite view lacks "reality" -- even to the point of being disconcerting. However -- and considering again that each and all views presented by an immersive video system in accordance with the present invention are drawn from the volume of a three-dimensional model - there is absolutely no reason that each and every view produced by an immersive video system should not be of absolute fidelity and correct spatial relationship to all other views.

For example, consider first the well known, but complex, pincushion correction circuitry of a common television. This circuitry serves to match the information modulation of the display-generating electron beam to the slightly non-planar, pincushion-like, surface of a common cathode ray tube. If the information extracted from a three-dimensional video model is so extracted in the contour of a common pincushion, then no correction of the information is required in presenting it on an equivalent pincushion surface of a cathode ray tube. Taking this analogy to the next level, if a scene is to be presented on some selected panels of a Liquid Crystal Digital (LCD) display, or walls of a room, then the pertinent video information as would constitute a perspective on the scene at each such panel or wall is simply withdrawn from the three-dimensional model. Because they are correctly spatially derived from a seamless 3-D model, the video presentations on each panel or wall fit together seamlessly, and perfectly.

By now, this capability of the immersive video of the present invention should be modestly interesting. As well as commonly lacking stereoscopy, the attenuation effects of intervening atmosphere, true color fidelity, and other assorted shortcomings, two-dimensional screen views of three-dimensional real world scenes suffer in realism because of subtle systematic dimensional distortion. The surface of the two-dimensional display screen (e.g., a television) is seldom so (optically) flat as is the surface of the Charge Coupled Device (CCD) of a camera providing a scene image. The immersive video system of the present invention straightens all this out, exactly matching (in dedicated embodiments) the image presented to the particular screen upon which the image is so presented. This is, of

course, a product of the 3-D video database which was itself constructed from multiple video streams from multiple video cameras. It might thus be said that the immersive video system of the present invention is using the image of one (or more) cameras to "correct" the presentation (not the imaging, the presentation) of an image derived (actually synthesized in part) from another camera!

2.3.2 Interactive Video Views on a Scene

Second, consider that immersive video in accordance with the present invention permits machine dynamic generation of views on a scene. Images of a real-world scene may be linked at the discretion of the viewer to any of a particular perspective on the scene, an object in the scene, or an event in the scene.

For example, consider again the example of the real-world event of an American football game. A viewer/user may interactively close to view a field goal attempt from the location of the goalpost crossbars (a perspective on the scene), watching a successful place kick sail overhead. The viewer/user may chose to have the football (an object in the scene) centered in a field of view that is 90° to the field of play (i.e., a perfect "sideline seat") at all times. Finally, the viewer/user may chose to view the scene from the position of the left shoulder of the defensive center linebacker unless the football is launched airborne (as a pass) (an event in the scene) from the offensive quarterback, in which case presentation reverts to broad angle aerial coverage of the secondary defensive backs.

The present and related inventions serve to make each and any viewer of a video or a television depicting a real-world scene to be his or her own proactive editor of the scene, having the ability to interactively dictate and select -- in advance of the unfolding of the scene, and by high-level command -- any reasonable parameter or perspective by which the scene will be depicted, as and when the scene unfolds.

2.3.3 Stereoscopic Video Views on a Scene

Third, consider that stereoscopy is inherent in immersive video in accordance with the present invention.

Scene views are constantly generated by reference to the content of a dynamic three-dimensional model -- which model is sort of a three-dimensional video memory without the storage requirement of a one-to-one correspondence between voxels (solid pixels) and memory storage addresses. Therefore it is "no effort at all" for an immersive video system to present, as a selected stream of video data containing a selected view, first scan time video data and second scan time video data that is displaced, each relative to the other, in accordance with the

location of each object depicted along the line of view.

This is, of course, the basis of stereoscopy. When one video stream is presented in a one color, or, more commonly at present, at a one time or in a one polarization, while the other video stream is presented in a separate color, or at a separate time, or in an orthogonal polarization, and each stream is separately gated to the eye (at greater than the eye flicker fusion frequency = 70 Hz) by action of colored glasses, or time-gated filters, or polarizing filters, then the image presented to the eyes will appear stereoscopic, and three-dimensional. The immersive video of the present invention, with its superior knowledge of the three-dimensional spatial positions of all objects in a scene, excels in such stereoscopic presentations (which stereoscopic presentations are, alas, impossible to show on he one-dimensional pages of the drawings).

2.3.4 A Combination of Visual Reality and Virtual Reality

Fourth, the immersive video presentations of the present invention are clearly susceptible of combination with the objects, characters and environments of artificial reality. Computer models and techniques for the generation and presentation of artificial reality commonly involve three-dimensional organization and processing, even if only for tracing light rays for both perspective and illumination. The central, "cartoon", characters and objects are often "finely wrought", and commonly appear visually pleasing. Alas, equal attention cannot be paid to each and every element of a scene, and the scene background to the focus characters and objects is often either stark, or unrealistic, or both.

Immersive video in accordance with the present invention provides the vast, relatively inexpensive, "database" of the real world (at all scales, time compressions/expansions, etc.) as a suitable "field of operation" (or "playground") for the characters of virtual reality.

When it is considered that immersive video permits viewer/user interactive viewing of a scene, then it is straightforward to understand that a viewer/use may "move" in and though a scene in response to what he/she "sees" in a composite scene of both a real, and an artificial virtual, nature. It is therefore possible, for example, to interactively flee from a "dinosaur" (a virtual animal) appearing in the scene of a real world city. It is therefore possible, for example, to strike a virtual "baseball" (a virtual object) appearing in the scene of a real world baseball park. It is therefore possible, for example, to watch a "tiger", or a "human actor" (both real animal) appearing in the scene of a virtual landscape (which landscape has been laid out in consideration of the movements of

the tiger or the actor).

Note that (i) visual reality and (ii) virtual reality can, in accordance with the present invention, be combined with (1) a synthesis of real/virtual video images/television pictures of a combination real-world/virtual scene wherein the synthesized pictures are to user-specified parameters of presentation, e.g. panoramic or at magnification if so desired by the user, and/or (2) the synthesis of said real/virtual video images/television pictures can be 3-D stereoscopic.

2.4 The Method of the Present Invention, In Brief

In brief, the present invention assumes, and uses, a three-dimensional model of the (i) static, and (ii) dynamic, environment of a real-world scene -- a three-dimensional, environmental, model.

Portions of each of multiple video streams showing a single scene, each from a different spatial perspective, that are identified to be (then, at the instant) static by a running comparison are "warped" onto the three-dimensional environmental model. This "warping" may be into 2-D (static) representations within the 3-D model -- e.g., a football field as is permanently static or even a football bench as is only normally static -- or, alternatively, as a reconstructed 3-D (static) object -- e.g., the goal posts.

The dynamic part of each video stream (that rises from a particular perspective) is likewise "warped" onto the three-dimensional environmental model. Normally the "warping" of dynamic objects is into a reconstructed three-dimensional (dynamic) objects -- e.g., a football player. This is for the simple reason that dynamic objects in the scene are of primary interest, and it is they that will later likely be important in synthesized views of the scene. However, the "warping" of a dynamic object may also be into a two-dimensional representation -- e.g., the stadium crowd producing a wave motion.

Simple changes in video data determine whether an object is (then) static or dynamic.

The environmental model itself determines whether any scene portion or scene object is to be warped onto itself as a two-dimensional representation or as a reconstructed three-dimensional object. The reason no attempt is made to reconstruct everything in three-dimensions are twofold. First, video data i slacking to model everything in and about the scene in three dimensions -- e.g., the underside of the field or the back of the crowd are not within any video stream. Second, and more importantly, there is insufficient computational power to reconstruct a three-dimensional video representation of everything that is within a scene, especially in real time

(i.e., as television).

Any desired scene view is then synthesized (alternatively, "extracted") from the representations and reconstituted objects that are (both) within the three-dimensional model, and is displayed to a viewer/user.

The synthesis/extraction may be in accordance with a viewer specified criterion, and may be dynamic in accordance with such criterion. For example, the viewer or a football game may request a consistent view from the "fifty yard line", or may alternatively ask to see all plays from the a stadium view at the line of scrimmage. The views presented may be dynamically selected in accordance with an object in the scene, or an event in the scene.

Any interior or exterior perspectives on the scene may be presented. For example, the viewer may request a view looking into a football game from the sideline position of a coach, or may request a view looking out of the football game from at the coach from the then position of the quarterback on the football field. Any requested view may be panoramic, or at any aspect ratio, in presentation. Views may also be magnified, or reduced in size.

Finally, any and all views can be rendered stereoscopically, as desired.

The synthesized/extracted video views may be processed in real time, as television.

Any and all synthesized/extracted video views contain only as much information as is within any of the multiple video streams; no video view can contain information that is not within any video stream, and will simply show black (or white) in this area.

2.5 The Immersive System of the Present Invention, In Brief

In brief, the immersive video computer system of the present invention receives multiple video images of view on a real world scene, and serves to synthesize a video image of the scene which synthesized image is not identical to any of the multiple received video images.

The computer system includes an information base containing a geometry of the real-world scene, shapes and dynamic behaviors expected from moving objects in the scene, plus, additionally, internal and external camera calibration models on the scene.

A video data analyzer means detects and tracks objects of potential interest in the scene, and the locations of these objects.

A three-dimensional environmental model builder records the detected and tracked objects at their proper locations in a three-dimensional model of the scene. This recording is in

consideration of the information base.

A viewer interface is responsive to a viewer of the scene to receive a viewer selection of a desired view on the scene. This selected and desired view need not be identical to any views that are within any of the multiple received video images.

Finally, a visualizer generates (alternatively, "synthesizes") (alternatively "extracts") from the three-dimensional model of the scene, and in accordance with the received desired view, a video image on the scene that so shows the scene from the desired view.

These and other aspects and attributes of the present invention will become increasingly clear upon reference to the following drawings and accompanying specification.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a top-level block diagram showing the high level architecture of the system for Multiple Perspective Interactive (MPI) video in accordance with the present invention.

Figure 2 is a functional block diagram showing an overview of the MPI system in accordance with the present invention, previously seen in block diagram in Figure 1, in use for interactive football video.

Figure 3 is a diagrammatic representation of the hardware configuration of the MPI system in accordance with the present invention, previously seen in block diagram in Figure 1.

Figure 4 is a pictorial representation of a video display particularly showing how, as a viewer interface feature of the Multiple Perspective Interactive (MPI) video system in accordance with the present invention previously seen in block diagram in Figure 1, a viewer can select one of the many items to focus in the scene.

Figure 5 is a diagrammatic representation showing how different cameras provide focus on different objects in the MPI system in accordance with the present invention; depending on the viewer's current interest an appropriate camera must be selected.

Figure 6 is another pictorial representation of the video display of the Multiple Perspective Interactive (MPI) video system in accordance with the present invention, this the video display particularly showing a viewer-controlled three-dimensional cursor serving to mark a point in three-dimensional (3-D) space, with the projection of the 3-D cursor being a regular 2-D cursor.

Figure 7 is a diagram showing coordinate systems for camera

calibration in the Multiple Perspective Interactive (MPI) video system in accordance with the present invention.

Figure 8, consisting of Figures 8a through 8c, is pictorial representation, and accompanying diagram, of three separate video displays in the Multiple Perspective Interactive (MPI) video system in accordance with the present invention, the three separate displays showing how three different cameras provide three different sequences, the three different sequences being used to build the model of events in the scene.

Figure 9, consisting of Figures 9a and 9b, is pictorial representation of two separate video displays in the Multiple Perspective Interactive (MPI) video system in accordance with the present invention showing many known points an image can be used for camera calibration; the frame of Figure 9a having sufficient points for calibration but the frame of Figure 9b having insufficient points for calibration.

Figure 10, consisting of Figures 10a through 10c, is pictorial representation of three separate video frames, arising from three separate algorithm-selected video cameras, in the Multiple Perspective Interactive (MPI) video system in accordance with the present invention.

Figure 11 is a schematic diagram showing a Global Multi-Perspective Perception System (GM-PPS) portion of the Multiple Perspective Interactive (MPI) video system in accordance with the present invention in use to take data from calibrated cameras covering a scene from different perspectives in order to dynamically detect, localize, track and model moving objects -- including a robot vehicle and human pedestrians -- in the scene.

Figure 12 is a top-level block diagram showing the high level architecture of the Global Multi-Perspective Perception System (GM-PPS) portion, previously seen in Figure 11, of the Multiple Perspective Interactive (MPI) video system in accordance with the present invention, the architecture showing the interaction between a priori information formalized in a static model and the information computed during system processing and used to formulate a dynamic model.

Figure 13 is a graphical illustration showing the intersection formed by the rectangular viewing frustum of each camera scene onto the environment volume in the GM-PPS portion of the MPI video system of the present invention; the filled frustum representing possible areas where the object can be located in the 3-D model while, by use of multiple views, the intersection of the frustum from each camera will closely approximate the 3-D location and form of the object in the environment model.

Figure 14, consisting of Figure 14a and Figure 14b, is a

diagram of a particular, exemplary, environment of use of the GM-PPS portion, and of the overall MPI video system of the present invention; the environment being an actual courtyard on the campus of the University of California, San Diego, where four cameras, the locations and optical axes of which are shown, monitor an environment consisting of static object, a moving robot vehicle, and several moving persons.

Figure 15 is a pictorial representation of the distributed architecture of the GM-PPS portion of the MPI video system of the present invention wherein (i) a graphics and visualization workstation acts as the modeler, (ii) several workstations on the network act as slaves which process individual frames based on the master's request so as to (iii) physically store the processed frames either locally, in a nearby storage server, or, in the real-time case, as digitized information on a local or nearby frame-grabber.

Figure 16 is a diagram showing the derivation of a camera coverage table for an area of interest, or environment, in which objects will be detected, localized, tracked and modeled by the GM-PPS portion of the MPI video system of the present invention; each grid cell in the area is associated with its image in each camera plane while, in addition, the diagram shows an object dynamically moving through the scene and the type of information the GM-PPS portion of the MPI video system uses to maintain knowledge about this object's identity.

Figure 17, consisting of Figures 17a through 17d, is four pictorial views of the campus courtyard previously diagrammed in Figure 14 at global time 00:22:29:06; the scene containing four moving objects including a vehicle, two walkers and a bicyclist.

Figure 18 is a pictorial view of a video display to the GM-PPS portion of the MPI video system of the present invention, the video display showing, as different components of the GM-PPS, views from the four cameras of Figure 17 in a top row, and a panoramic view of the model showing hypotheses corresponding to the four moving objects in the scene in a bottom portion; the GM-PPS serving to detect each object in one or more views as is particularly shown by the bounding boxes, and serving to update object hypotheses by a line-of-sight projection of each observation.

Figure 19, consisting of Figures 19a through 19e, is five pictorial views of the GM-PPS model showing various hypotheses corresponding to the four moving objects in the scene of Figure 17 at global time 00:22:29:06; Figures 19a-19d correspond to four actual camera views while Figure 19e shows a virtual image from the top of the scene.

Figure 20, consisting of Figures 20a through 20d, is four pictorial views of the same campus courtyard previously

diagrammed in Figure 14, and shown in Figure 17, at global time 00:62:39:06; the scene still containing four moving objects including a vehicle, two walkers and a bicyclist.

Figure 21 is another pictorial view of the video display to the GM-PPS portion of the MPI video system of the present invention previously seen in Figure 18, the video display now showing a panoramic view of the model showing the hypotheses corresponding to the four moving objects in the scene at the global time 00:22:39:06 as was previously shown in Figure 20.

Figure 22, consisting of Figures 22a through 22c, is a diagrammatic view showing how immersive video in accordance with the present invention uses video streams from multiple strategically-located cameras that monitor a real-world scene from different spatial perspectives.

Figure 23 is a schematic block diagram of the software architecture of the immersive video system in accordance with the present invention.

Figure 24 is a pictorial view showing how the video data analyzer portion of the immersive video system of the present invention detects and tracks objects of potential interest and their locations in the scene.

Figure 25 is a diagrammatic view showing how, in an immersive video system in accordance with the present invention, the three-dimensional (3D) shapes of all moving objects are found by intersecting the viewing frustrums of objects found by the video data analyzer; two views of a full three-dimensional model generated by the environmental model builder of the immersive video system of the present invention for an indoor karate demonstration being particularly shown.

Figure 26 is a pictorial view showing how, in the immersive video system in accordance with the present invention, a remote viewer is able to walk though, and observe a scene from anywhere using virtual reality control devices such as the boom shown here.

Figure 27 is an original video frame showing video views from four cameras simultaneously recording the scene of a campus courtyard at a particular instant of time.

Figure 28 is four selected virtual camera, or synthetic video, images taken from a 116-frame "walk through" sequence generated by the immersive video system in accordance with the present invention (color differences in the original color video are lost in monochrome illustration).

Figure 29, consisting of Figures 29a and Figure 29b, are synthetic video images generated from original video by the immersive video system in accordance with the present invention, the synthetic images respectively showing a "bird's eye view" and a ground level view of the same courtyard previously seen in

Figure 27 at the same instant of time.

Figure 30a is a graphical rendition of the 3D environment model generated for the same time instant shown in Figure 27, the volume of voxels in the model intentionally being at a scale sufficiently coarse so that the 3D environmental model of two humans appearing in the scene may be recognized without being so fine that it cannot be recognized that it is only a 3D model, and not an image, that is depicted.

Figure 30b is a graphical rendition of the full 3D environment model generated by the environmental model builder of the immersive video system of the present invention for an indoor karate demonstration as was previously shown in Figure 25, the two human participants being clothed in karate clothing with a kick in progress, the scale and the resolution of the model being clearly observable.

Figure 30c is another graphical rendition of the full 3D environment model generated by the environmental model builder of the immersive video system of the present invention, this time for an outdoor karate demonstration, this time the environmental model being further shown to be located in the static scene, particularly of an outdoor courtyard.

Figure 31 is a listing of Algorithm 1, the Vista "Compositing" or "Hypermosaicing" Algorithm, in accompaniment to a diagrammatic representation of the terms of the algorithm, of the present invention where, at each time instant, multiple vistas are computed using the current dynamic model and video streams from multiple perspective; for stereoscopic presentations vistas are created from left and from right cameras.

Figure 32 is a listing of Algorithm 2, the Voxel Construction and Visualization for Moving Objects Algorithm in accordance with the present invention.

Figure 33 is a synthetic video frames, similar to the frames of Figure 10, created by the immersive video system of the present invention at a random user-specified viewpoint during a performance of a indoor karate exercise by an actor in the scene, the virtual views of an indoor karate exercise of Figure 33 being rendered at a higher resolution than were the virtual views of the outdoor karate exercise of Figure 30.

DESCRIPTION OF THE PREFERRED EMBODIMENT

1. Capabilities of the Multiple Perspective Interactive Video of the Present Invention, and Certain Potential

Implications of These Capabilities

The capabilities of the Multiple Perspective Interactive (MPI) video of the present invention are discussed even prior to teaching the system that realizes these capabilities in order that certain potential implications of these capabilities may best be understood. Should these implications be understood, it may soon be recognized that the present invention accords not merely a "fancy form" of video, but an in-depth change to the existing, fundamental, video and television viewing experience.

The present specification presents a system, a method and a model for Multiple Perspective Interactive -- "MPI" -- video or television. In the MPI video model multiple cameras are used to acquire an episode or a program of interest from several different spatial perspectives. The cameras are real, and exist in the real world: to use a source camera, or a source image, that is itself virtual constitutes a second-level extension of the invention, and is not presently contemplated.

MPI video is always interactive -- the "I" in MPI -- in the sense that the perspective from which the video scene is desired to be, and will be, shown and presented to a viewer is permissively chosen by such viewer, and predetermined. However, MPI video is also interactive in that, quite commonly, the perspective on the scene is dynamic, and responsive to developments in the scene. This may be the case regardless that the real video images of the scene from which the MPI video is formed are themselves dynamic and may, for example, exhibit pan and zoom. Accordingly, a viewer-selected dynamic presentation of dynamic events that are themselves dynamically imaged is contemplated by the present invention.

Consider, for example, the presentation of MPI video for a game of American football. The "viewer-selected dynamic presentation" might be, for example, a viewer-selected imaging of the quarterback. This image is dynamic in accordance that the quarterback should, by his movement during play, cause that, in the simplest case, the images of several different video camera should be successively selected or, in the case of such full virtual video as is contemplated by the present invention, that the quarterback's image should be variously dynamically synthesized by digital computer means. The football game is, of course, a dynamic event wherein the quarterback moves. Finally, the real-world source, camera, images that are used to produce the MPI video are themselves dynamic in accordance that the cameramen at the football game attempt to follow play.

The net effect of all this dynamism is non-obvious, and of a different order than even such video, or television, experience as is commonly accorded a network video director of a

major sporting event who is exposed to a multitude of (live) video feeds. The experience of MPI video in accordance with the present invention may usefully be compared, and contrasted, with virtual reality. The term "virtual reality" commonly has connotations of (i) unreality, (ii) sensory immersion, and/or (iii) self-directed interaction with a reality that is only fantasy, or "virtual". The effect of the MPI video of the present invention differs from &virtual reality" in all these factors, but is nonetheless quite shocking.

In the first place, the present invention is not restricted to use with video depicting reality -- but reality is the cheapest source of such information as can, when viewed through the MPI video system of the present invention, still be quite "intense". In other words, it may be necessary to be attacked by a fake, virtual, tiger when one can visually experience the onrush of a real hostile football linebacker.

In the second place, MPI video is presented upon a common monitor, or television set, and does not induce the viewer to believe that he or she has entered a fantasy reality.

Finally, and in the third place, the self-directed interaction with MPI video is directed to observational perspective, and not to a viewer's dynamic control of developments in the scene in accordance with his or her action, or inaction.

What MPI video can do, and what causes it to be "shocking", is that the viewer can view, or, in the American vernacular, "get into", the video scene just where, and even when, the viewer chooses. Who at a live sporting event has not looked at the cheerleaders, a favorite player, or even the referee? Psychological and sociological research has shown that, among numerous other differences between us all, men and women, as one example, do not invariably visually acquire the same elements of a picture or painting, let alone do the two sexes visually linger on such elements as they identify in common for equal time durations. (Women like to look at babies in a scene more so than do men, and men like to look at women in a scene more so than do women.) Quite simply, humans often have different interests, and focal points of interest, even in the same visual subject matter. With present video and television presentations everyone must watch the same thing, a "common composite". With the viewer-interactive control that is inherent in MPI video, different things can be differently regarded at each viewer's behest. Accordingly, MPI video removes some of the limitations that presently make a video or a television viewer only a passive participant in the video or television viewing process (in the American vernacular, a "couch potato").

Of course, MPI video need not be implemented for each and

every individual video or television viewer in order to be useful. Perhaps with the advent of communicating 500 channels of television to the home, a broadcast major American football game might reasonably consume not one, but 25+ channels -- one for each player of both sides on the football field, one for each coach, one for the football, and one for the stadium, etc.

En early alternative may be MPI video on pay per view. It has been hypothesized that the Internet, in particular, may expand in the future to as likely connect smart machines to human users, and to each other, as to communicatively interconnect more and more humans, only. Customized remote viewing can certainly be obtained by assigning every one his or her own remotely-controllable TV camera, and robotic rover. However, this scheme soon breaks down. How can hundreds and thousands of individually-remotely-controlled cameras jockey for position and for viewer-desired vantage points at a single event, such as the birth of a whale, or an auto race? It is likely a better idea to construct a comprehensive video image database from quality images obtained from only a few strategically positioned cameras, and to then permit universal construction of customized views from this database, all as is taught by the present invention.

As will additionally be seen, the MPI video of the present invention causes video databases to be built in which databases are contained -- dynamically and from moment to moment (frame to frame) -- much useful information that is interpretive of the scene depicted. Clearly, in order to select, or to synthesize, an image of a particular player, the MPI video system contains information of the player's present whereabouts, and image. It is thus a straightforward matter for the system to provide information, in the form of text or otherwise, on the scene viewed, either continuously or upon request.

Such auxiliary information can augment the entertainment experience. For example, a viewer might be alerted to a changed association of a football in motion from a member of a one team to a member of the opposing team as is recognized by the system to be a fumble recovery or interception. For example, a viewer might simply be kept informed as to which player presently has possession of the football.

The more probable use of such auxiliary information is education. It will no longer be necessary to remain in confused ignorance of what one is viewing if, by certain simple commands, "helps" to understanding the scene, and the experience, may be obtained.

2. An Actual System Performing Multiple Perspective Interactive (MPI) Video in Accordance With the Present

Invention, and Certain Limitations of this Exemplary System

The MPI video model, its implementation, and the architectural components of a rudimentary system implementing the model are taught in the following sections 3 through __ of this specification. Television is a real-time version of MPI video. Interactive TV is a special case of MPI video. In MPI TV, many operations must be done in real time because many television programs are broadcast in real time.

The concept of MPI video is taught in the context of a sport event. The MPI video model allows a viewer to be active; he or she may request a preferred camera position or angle, or the viewer may even ask questions about contents described in the video. Even the rudimentary system automatically determines the best camera and view to satisfy the demands of the viewer.

Videos of American football have been selected as the video source texts upon which the performance of MPI video will be taught and demonstrated. Football video already in existence was retrieved, and operated upon as a sample application of MPI video in order to demonstrate certain desirable features.

The particular, rudimentary, embodiment of an MPI video system features automatic camera selection and interaction using three-dimensional cursers. The complete computational techniques used in the rudimentary system are not fully contained herein this specification in detail because, by an large, know techniques hereinafter referred to are implemented. Certain computational techniques are, however, believed novel, and the mathematical basis of each of these few techniques are fully explained herein.

The rudimentary, demonstration, system of the present invention has been reduced to operative practice, and ell drawings or photographs of the present specification that appear to be of video screens are representations or photographs of actual screens, and are not mock-ups. Additionally, where continuity between successive video views is implied, then this continuity exists in reality although, commensurate with the amount of computer resource and computational power harnessed to do the necessary transformations, the successive and continuous views and presentations may not be in full real time.

The running MPI video system is presently being extend to other applications besides American football. In particular, a detail teaching of the concept, and method, of generating a three-dimensional database required by the MPI video system of the present invention is taught and demonstrated in this specification not in the context of football, but rather, as a useful simplification, in the context of a university courtyard though which human and machine subjects (as opposed to football players) roam. The present specification will accordingly be

understood as being directed to the enabling principles, construction, features and resulting performance of rudimentary embodiment of an MPI video system, as opposed to presenting great details on any or all of the several separate aspects of the system.

3. Architecture of the MPI Video System

A physical phenomena or an event can be usually viewed from multiple perspectives. The ability to view from multiple perspectives is essential in many applications. Current remote viewing via video or television permits viewing only from one perspective, and that perspective being that of an author or editor and not of the viewer. A viewer has no choice. However, remote viewing via video or television even under such limitations has been very attractive and has influenced our modern society in many aspects.

Technology has now advanced to the state that each of many simultaneous remote viewers (i) can be provided with a choice to so view remotely from whatever perspective they want, and, with limitations, (ii) can interactively select just what in the remote scene they want to view.

Let us assume that an episode is being recorded, or being viewed in real time. This episode could be related, for example, to a scientific experiment, an engineering analysis, a security post, a sports event, or a movie. In a simplest and most obvious case, the episode can be recorded using multiple cameras strategically located at different points. These cameras provide different perspectives of the episode. Each camera view is individually very limited. The famous parable about an elephant and the blind men may be recalled. With just one camera, only a narrow aspect of the episode may be viewed. Like a single blind man, a single camera is unable to provide a global description of an episode.

Using computer vision and related techniques in accordance with the present invention, it is possible to take individual camera views and reconstruct an entire scene. These individual camera scenes are then assimilated into a model that represents the complete episode. This model is called an "environment model". The environment model has a global view of the episode, and it also knows where each individual camera is. The environment model is used in the MPI system to permit a viewer to view what he or she wants from where he or she wants (within the scene, and within limits).

Assume that a viewer is interested in one of the following.

First, the viewer may be interested in a specific perspective, and may want to view a scene, an episode, or an entire video presentation from this specific perspective. The

user may specify a real, or a virtual, camera specifically. Alternatively, the viewer may only specify the desired general location of the camera, without actual knowledge whether a camera in such location would be real or virtual.

Second, the viewer may be interested in a specific object. There may be several objects in a scene, an episode, or a presentation. A viewer may want to always view a particular object independent of its situation in the scene, episode, or presentation. Alternatively, the object that is desired to be viewed may be context sensitive: the viewer may desire view the basketball until the goal is scored to then shift view to the last player to touch the basketball.

Third, the viewer may be interested in a specific event. A viewer may specify characteristics of an event and may want to view a scene, an episode, or a presentation from the best perspective for that event.

Fourth, the viewer may be interested in a having a view from a virtual camera. The viewer may request to view a scene of an event within the scene from a perspective that is not provided by any real camera that is situated to acquire the scene or any portion thereof. In such cases, the MPI video system of the present invention will, by use of the environment model and video synthesis techniques, synthesize a virtual camera, and video image, so as to view a scene, an episode, or an entire presentation from a viewer-specified perspective.

The high level architecture for a MPI video system so functioning is shown in a first level block diagram in Figure 1. A image at a certain perspective from each camera 10a, 10b, ...10n is converted to its associated camera scene in camera screen buffers CSB 11a, 11b, ...11n. Multiple camera scenes are then assimilated into the environment model 13 by computer process in the Environ. Model Builder 12. A viewer 14 (shown in phantom line for not being part of the MPI video system of the present invention) can select his perspective at the Viewer Interface 15, and that perspective is communicated to the Environment Model via a computer process in Query Generator 16. The programmed reasoning system in the Environment Model 13 decides what to send via Display Control 17 to the Display 18 of the viewer 14.

Implementation of a universal, plug and play, MPI video system that (i) track virtually anything, (ii) function in real time (i.e, for television), and/or (iii) produce virtually any desired image, including a full virtual image, severely stresses modern computer and video hardware technology circa 1995, and can quickly come to consume the processing power of a mini-supercomputer. Economical deployment of the MPI video system requires, circa 1995, advances in several hardware technology

areas. Notably, however, there is, as will imminently be demonstrated, no basic hardware nor software function required by such a MPI video system that ms not only presently realizable, but that is, in actual fact, already realized. Moreover, a relatively high level, user friendly, viewer interface -- which might have been considered impossible or extremely difficult of being successfully achieved -- "falls out" quite naturally, and to good effect, from the preferred implementation of, and the partitioning of function within, the MPI system.

A complete MPI video system with limited features can be, and has been, implemented using the existing technology. The exact preferred architecture of a MPI video system will depend on the area to which the system is intended to be applied, and the type and level of viewer interaction allowed. However, certain general issues are in common to any and all implementations of MPI video systems. Seven critical areas that must be addressed in building any MPI video system are as follows.

First, a camera scene builder is required as a programmed computer process. In order to convert an image sequence of a camera to a scene sequence, the MPI video system must, and does, know where the camera is located, its orientation, and its lens parameters. Using this information, the MPI video system is then able to locate objects of potential interest, and the locations of these objects in the scene. This requires powerful image segmentation methods. For structured applications, the MPI video system may use some knowledge of the domain, and may even change or label objects to make its segmentation task easier. This is, in fact, the approach of the rudimentary embodiment of the MPI video system, as will be further discussed later.

Second, an environment model builder is required as a programmed computer process. Individual camera scenes are combined in the MPI video system to form a model of the environment. All potential objects of interest and their locations are recorded in the environment model. The representation of the environment model depend on the facilities provided to the viewer. If the images are segmented properly, then, by use of powerful but known computers and computing methods, it is possible to build environment models in real time, or almost in real time.

Third, a viewer interface permits the viewer to select the perspective that he or she wants. This information is obtained from the user in a friendly but directed manner. Adequate tools are provided to the user to point and to pick objects of interest, to select the desired perspective, and to specify

SUBSTITUTE SHEET (RULE 26)

events of interest. Recent advances in visual interfaces, virtual reality, and related areas have contributed to making the MPI video system viewer interface very powerful -- even in the rudimentary embodiment of the system.

Fourth, a display controller software process is required to respond to the viewers' requests by selecting appropriate images to be displayed to each such viewer. These images may all come from one perspective, or the MPI video system may have to select the best camera at every point in time in order to display the selected view and perspective. Accordingly, multiple cameras may be used to display a sequence over time, but at any given time only a single best camera is used. This has required solving a camera hand-off problem.

Fifth, a video database must be maintained. If the video event is not in real time (i.e., television) then, then it is possible to store an entire episode in a video database. Each camera sequence is stored along with its metadata. Some of the metadata is feature based, and permits content-based operations. See Ramesh Jain and Arun Hampapur; "Metadata for video-databases" appearing in *SIGMOD Records*, Dec. 1994.

In many applications of the MPI video system, environment models are also stored in the database to allow rapid interactions with the system.

Sixth, real-time processing of video must be implemented to permit viewing of real time video events, i.e. television. In this case a special system architecture is required to interpret each camera sequence in real time and to assimilate their results in real time so that, based on a viewer input, the MPI video system can use the environment model to solve the camera selection problem.

A practitioner of the computer arts and sciences will recognize that this sixth requirement is nothing but the fifth requirement performed faster, and in real time. The requirement might just barely be realizable in software if computational parallelism is exploited, but, depending upon simplifying assumptions made, a computer ranging from an engineering work station to a full-blown supercomputer (both circa 1995) may be required. Luckily, low-cost (but powerful) microprocessors are likely distributable to each of the Camera Sequence Buffers CSB 11a, 11b, ...11n in order to isolate, and to report, features and dynamic features within each camera scene. Correlation of scene features at a higher process level may thus be reduced to a tractable problem. Another excellent way of simplifying the problem -- which way is used in the rudimentary embodiment of the MPI video system taught within this specification -- is to demand that the scene, and each camera view thereof, include constant, and readily identifiable, markers as a sort of video

"grid". An American football field already has this grid in the form of yard lines and hash marks. So might a college courtyard with benches and trees. A whale free swimming in an amorphous tank while giving birth is at the other end of the spectrum, and presents an exceedingly severe camera image selection (if not also correlation) problem.

Seventh, a visualizer is required in those applications that require the displaying of a synthetic image in order to satisfy a viewer's request. For example, it is possible that a user selects a perspective that is not available from any camera. A trivial solution is simply to select the closest camera, and to use its image. The solution of the rudimentary MPI video system of the present specification -- which solution is far from trivial in implementation or trite in the benefits obtained -- is to select a best -- and not necessarily a closet -- camera and to use its image and sequence.

The ultimate response of the MPI video system is to synthesize the exact synthetic image, and image sequence, the viewer desires and demands. Even here, no image can be formed where no source image data exists, such as a view from below the playing field (i.e., from in the ground). Even a synthetic view that is normally acceptable, such as "from the nose of the football in the vector direction of the movement of same" cannot be produced when, and at such times as, the football becomes "buried", and obscured from view, under a pile up after the ball carrier is tackled. "Weird" views in synthesized MPI video can be exciting, but, in accordance with their "weirdness", are not always reliably capable of being successfully synthesized.

The ability of an MPI video system to synthesize a full virtual video image is basically a function of "raw" computational power. If real time video (i.e., television) is not required, short virtual video segments of real world occurrences may be quite as reasonably produced, and maybe more reasonably produced, than the computer-generated special effects, including morphing, so popular in American movies circa 1995. Of course, it should be understood that even the synthesis of such segments requires computers of considerable speed capacity.

Clearly, implementation of an MPI video system with unrestricted capability requires state-of-the art computer hardware and software, and will benefit by such improvement in both as are confidently expected. Some new issues, other than the above seven, are expected to arise in addressing different applications of MPI video. At the present time, and in this specification, only a rudimentary MPI vide system is taught. By implementing this first MPI video system, the inventors have identified interesting future issues in each of computer vision,

artificial intelligence, human interfaces, and databases. However, and for the moment, the following sections serve to discuss and teach an actual MPI video system that was implemented to demonstrate the concept of the invention more concretely and completely, as well as to define and identify performance issues.

4. A Rudimentary, Prototype, Embodiment of an MPI Video System in Use for Producing MPI Video of American Football

Key concepts in MPI video are taught in this section 4. by reference to a rudimentary, prototype, embodiment of an MPI video system that was built particularly for multiple perspective interactive viewing of American football. The motivation of the inventors in selecting this domain was to find a domain that was realistic, interesting, non-trivial and sufficiently well structured so as to demonstrate many important concepts of MPI video. It is also of note that, should the present MPI video system be applied commercially, it might already be possessed of such characteristics as would seemingly make it of some practical use in certain applications such as the "instant replay".

Many other sports and many other applications were considered by the inventors. American football was chosen due to the several attributes of the game that make it highly structured both from (i) database and (ii) computer vision perspective. These issues of structure are hereinafter discussed in the context of the implementation of the rudimentary, prototype, embodiment of the MPI video system

4.1 Scenario of Use, and Required Functions, of an MPI Video System As Applied to American Football

Although American-type football games are very popular in North America on conventional television, the broadcasts of these football games have several limitations from a viewer's perspective. The viewing of American football games could seemingly be significantly enhanced by adding the following facilities.

Usually a football game is captured by several cameras that are placed at different locations on the field. Though those cameras cover various parts of the game, viewers can get only one camera view at a time. This view is not a result of viewers' choice, but is instead what an editor thinks most people want to see. In most cases, editor's decision are right. In any case, with the current technology this expert selection of views is seemingly the best that can be done. If a viewer is interested in a certain player, or a shot from a different angle, than he or she cannot see the desired image unless the

editor's choice happens to be the same as the viewer's. By giving choices to a viewer, it is anticipated that watching the game might be made significantly more interesting.

Moreover, when watching football game questions often occur to viewers such as "who is this player who just now tackled", or "how long did this player run in this play". Conventional video or television does not necessary provide such information. Tools that provide such information would seemingly be useful.

Still further, while watching a video of a football game, a coach or a player may want to analyze how a particular player ran, or tackled, and to ignore all other players. An interactive viewing system should allow the viewing of only plays of interest, and these from different angles. Moreover, the video would desirably be good enough so that some detailed analysis would be capable of being performed on the video of the plays in order to study the precise patterns, and performance, of the selected player.

In the rudimentary MPI video system, viewers may both (i) select cameras according to their preference, and (ii) ask questions about the name(s), or the movement(s), of players. The following are some examples of interaction between a viewers and the MPI video system

The viewer may request that the MPI video system should show a shot of some upcoming play or plays taken from camera located behind the quarterback.

The viewer may request that the MPI video system should show a best shot of a particular, viewer-identified, player.

The viewer may request that the MPI video system should show as text the name of the player to which the viewer points, with his or her cursor, on the screen of the display 18 (shown in Figure 1).

The viewer may request that the MPI video system should highlight on the screen a particular player whose name the viewer has selected from a player list.

The viewer may request that the MPI video system should show him or her the exact present location of a selected player.

The viewer may request that the MPI video system should show him or her the sequence when a selected player crossed, for example, the 40 yard line.

The viewer may request that the MPI video system should show him or her the event of a fumble.

The viewer may request that the MPI video system should show all third down plays in which quarterback X threw the ball to the receiver Y.

To perform these functions, and others, the MPI video system needs to have information about (i) contents of the football scene as well as (ii) video data.

Some of the above, and several similar questions, are relevant to MPI television, while others are relevant to MPI video. The major distinction between MPI TV and MPI video is in the role of the database. In case of MPI video, it is assume that much preprocessing can transpire, with the pre-processed information stored in a database. In case of MPI TV, most processing must be, and will be, in real time.

In the following section the rudimentary, prototype, MPI system discussed is, remarkably, an MPI TV system. A large random access video database system that is usable as an component of an MPI video system is realizable by conventional means, but is expensive (circa 1995) in accordance with amount of video stored, and the rapidity of the retrieval thereof.

In the rudimentary, prototype, MPI TV system, as shown in Fig. 2, a football scene is captured by several cameras and analyzed by a scene analysis system. The information obtained from individual cameras is used to form the environment model. The environment model allows viewers to interactively view the scene.

Additionally, a prototype football video retrieval system haw been implemented, as hereafter explained, This system incorporates some of the above-listed functions such as automatic camera selection and pointing to players. Other functions are readily susceptible of implementation using the same, existing, hardware and software technologies as are already within the rudimentary embodiment of the system.

4.1.1 Overview of the MPI Football Video/Television System

The configuration of the MPI football video/television system is shown in Figure 3. The current system consists of a UNIX workstation, a laser disc player, a video capture board, and a TV monitor and graphical display. The TV monitor is connected to the laser disc player. The laser disc player is controlled by the UNIX workstation. A graphical user interface is built using X-window and Motif on graphical display.

In use of the system, video of a football game was recorded on a laser disc. The actual video recorded was a part of the 1994 Super Bowl game. Since this video footage was obtained by commercial broadcast, the inventors did not have any control on camera location. Instead, the camera positions were reverse engineered using camera calibration algorithms. See R. M. Haralick and L. G. Shapiro; *Computer and Robot Vision*, Addison-Wesley Publishing, 1993.

Next, selected parts of the Super Bowl football game in which views from three different cameras were shown were selected. The three views were, of course, broadcast at three separate times. They depict an important, and exciting, play in

the 1994 Super Bowl game. This selection was necessary to simulate the availability of separate video streams from multiple cameras.

This video data was divided into shots, each of which corresponds to one football play. Each shot was analyzed and a three-dimensional scene description -- to be discussed in considerable detail in sections 5-10 hereinafter -- was generated. Shots from multiple cameras were combined into the environment model. The environment model contains information about position of players and status of cameras. The environment model is used by the system to allow MPI video viewing to a user. User commands are treated as queries to the system and are handled by the environment model and the database.

The interactive video interface of the system is shown in Figure 4. The video screen of Figure 4 shows video frames taken from laser disc. Video control buttons control video playback. Using a camera list, a viewer can choose any camera. Using a player list, a viewer can choose certain players to be focused on. If a viewer doesn't select a camera, then the system automatically selects the best camera. Also, multiple viewers can interact using the three-dimensional cursor. These new features are described below. Some interface features for the interactive video are shown here. A user can select one of the many items to focus in the scene.

4.2 Automatic Camera Selection

At any moment, there are several cameras that shoot the game. Automatic camera selection is a function that selects the best camera according to the preference of a user. Suppose a player is captured by three cameras and they produce three views shown in Fig. 5. In this case camera 2 is the best to see this player, for in camera 1 the player is out of the area while in camera 3 the player is too small. Different cameras provide focus on different objects. Depending on the current interest, an appropriate camera must be selected.

This function is performed by the system in the following way. First, viewers select the player that they want to see. Then the system looks into information on player position and camera status in the environment model to determine which camera provides the best shot of the player. Finally the selected shot is routed to the screen.

4.3 Interaction Using Three-Dimensional Cursers

In accordance with the present invention, a three-dimensional cursor is introduced in support of the interaction between viewers and the MPI vide/TV system. A

three-dimensional cursor is a cursor that moves in three-dimensional space. It is used to indicate particular position in the scene. The MPI video/TV system uses this cursor to highlight players. Viewers also use it to specify players that they want to ask questions about.

Examples of interaction using three-dimensional cursers are shown in Figure 6. As shown in Figure 6, the cursor consists of five lines. Three of the five lines indicate the x, y and z axes of the three-dimensional space. The intersection of these three lines shows cursor position. The other two lines indicate a projection of the three lines onto the ground. The projection helps viewers have a correct information of cursor position.

A viewer can manipulate the three-dimensional cursor so as to mark a point in the three-dimensional space. The projection of the three dimensional cursor is a regular cursor centered at the projection of this marked point.

Both viewers and the MPI system use the three-dimensional cursor to interact with each other. In the first example of Figure 6, a viewer moves the cursor to the position of a player and asks who this player is. The MPI system then compares the position of the cursor and the present position of each player to determine which player the viewer is pointing.

In the second example of Fig. 6, a viewer tells the MPI system a name of a player and asks where the player is. The MPI system then shows the picture of the player and overlays the cursor on the position of the player so as to highlight the player.

5. Three-dimensional Scene Analysis

The purpose of scene analysis is to extract three-dimensional information from video frames captured by cameras. This process is performed in the following two stages:

First, 2-D information is extracted. From each video frame, feature points such as players and field marks ere extracted and a list of feature points is generated.

Second, 3-D information is extracted. From the two-dimensional description of the video frame, three-dimensional information in the scene, such as player position and camera status, is then extracted.

The details of these extractions are contained within the following sub-sections.

5.1 Extracting Two-dimensional Information

In the extraction of two-dimensional information, feature points are extracted from each video frame. Feature points include two separate items in the images. First, the players are defined by using their feet as feature points. Second, the

field marks of the football field are sued as feature points. As is known to fans of American football, and American football field has yard lines to indicate yardage between goal lines, and hash marks to indicate a set distance from the side border, or sidelines, of the field. Field marks are defined as feature points because their exact position as a prior known, and their registration and detection can be used to determine camera status.

In the rudimentary, prototype, MPI system, the feature points are extracted by human-machine interaction. This process is currently carried out as follows. First, the system displays a video frame on the screen of Display 18 (shown in Figure 1). A viewer, or operator, 14 locates some feature points on the screen and inputs required information for each feature point. The system reads image coordinates of the feature points and generates two-dimensional description.

This process results in two-dimensional description of a video frame that consists of a list describing the players and a list describing the field marks. The player descriptions include each player's name and the coordinates of each player's image. The field mark descriptions include the positions (in the three-dimensional world), and the image coordinates, of all the field marks.

In the rudimentary embodiment of the MPI video system, all feature points are specified interactively with the aid of human intelligence. Many features can be detected automatically using machine vision techniques. See R. M. Haralick and L. G. Shapiro, op cit. The process of automatically detecting features in arbitrary images is not trivial, however. It is anticipated, however, that two trends will help the process of feature point identification in MPI video. First, new techniques have recently been developed, and will likely continue to be developed, that should be useful in permitting the MPI video system to extract feature point information automatically. Future new techniques may include some bar-code like mechanism for each player, fluorescent coloring on the players' helmets, or even some simple active devices that will automatically provide the location of each player to the system. It is also anticipated that many current techniques for dynamic vision and related areas may suitably be adapted for the MPI video application.

Because the goal of the rudimentary, prototype, system is primarily to demonstrate MPI video, no extensive effort has been made to extract the feature points automatically. Further progress, and greater system capabilities, in this area is deemed straightforward, and susceptible of implementation by a practitioner of the digital video

5.2 Extracting Three-dimensional Information

The purpose of this step is to obtain three-dimensional information from the two-dimensional frames. The spatial relationship between the three-dimensional world and the video frames captured by the cameras is shown in Figure 7. Consider that a camera is observing a point (x, y, z). A point (u, v) in the image coordinate system to which the point (x, y, z) is mapped may be determined by the following relationships, which relationships comprise a coordinate system for camera calibration.

A point (x, y, z) in the world coordinate system is transformed to a point (p, q, s) in the camera coordinate system by the following equation

$$\begin{bmatrix} p \\ q \\ s \end{bmatrix} = R \begin{bmatrix} x - x_0 \\ y - y_0 \\ z - z_0 \end{bmatrix}$$

where R is a transformation matrix from the world coordinate system to the camera coordinate system, and (x_0, y_0, z_0) is the position of the camera.

A point (p,q,s) in the camera coordinate system is projected to point (u,v) on the image plane according to the following equation:

$$\begin{bmatrix} u \\ v \end{bmatrix} = f / s \begin{bmatrix} p \\ q \end{bmatrix}$$

where f is camera parameter that determines the degree of zoom in or zoom out.

Thus, we see that an image coordinate (u,v) which corresponds to world coordinate (x,y,z) is determined depending on the (i) camera position, (ii) camera angle and (ii) camera parameter.

Therefore, from two-dimensional information that is described above, we can obtain three-dimensional camera and player information in the following way. (See R. M. Haralick and L. G. Shapiro; *Computer and Robot Vision*, Addison-Wesley Publishing, 1993.)

First, a camera calibration is performed. If only one known point is observed, a pair of image coordinates and world coordinates may be known. By applying this known pair to the above equations, two equations regarding the seven parameters that determine camera status may be obtained. Observing at

least four known points will suffice to provide the minimum equations to solve the seven unknown parameters.

However in the application of the MPI video system to football, the (i) camera position is usually fixed, and (ii) the rotation angle is zero. This reduces the number of unknowns to three, which requires minimum of two known points. The field marks extracted in previous process are then used as known points.

Next, an image to world coordinate mapping is performed. Once the camera status -- which is described by the seven parameters above -- is known, the world coordinate may be determined from the image coordinate if it considered that the point is constrained to lie in a plane. In the application of the MPI video system to football, the imaged football players are always approximately on the ground. Accordingly, the positions of players can be determined according to the above equations.

5.3 Interpolation

Ideally the scene analysis process just described should be applied to every video frame in order to get the most precise information about (i) the location of players and (ii) the events in the scene. However, it would require significant human and computational effort to do so in the rudimentary, prototype, MPI video system because feature points are located manually, end not by automation. Therefore, one key frame has been manually selected for every thirty frames, and scene analysis has been applied to the selected key frames. For frames in between, player position and camera status is estimated by interpolation between key frames by proceeding under the assumption that coordinate values change linearly between consecutive two key frames.

5.4 Camera Hand-Off

The rudimentary, prototype, MPI video system is able to determine and select a single best camera to show a particular player or an event. This is determined by the system using the environment model. Effectively, for the given player's location, the system uses reverse mapping for given camera locations, and then determines where will the image of the player be in the image for different cameras.

At the present time, the system selects the camera in which the selected player is closest to the center of the viewing area. The system could prospectively be made more precise by considering the orientation of the player also. The problem of transferring display control from one camera to another is called the "camera hand-off problem".

6. Results of the Exercise of the Rudimentary MPI Video System

The rudimentary, prototype, MPI video system has been exercised on a very simple football scene imaged from three different cameras. The goal of this example is to demonstrate the method and apparatus of the invention, and the feasibility of obtaining practical results. The present implementation and embodiment can clearly be extended process longer sequences, and also to different applications, and, indeed, is already being so extended.

The actual video data used in the experimental exercise of the MPI video system is shown in Figure 8. The video data consists of the three shots respectively shown in Figures 8a through 8c. These three shots record the same football play but are taken from different camera angles. Each shot lasted about ten seconds. The three different cameras thus provide three separate, but related, sequences. These sequences are used to build the model of events in the scene.

Key frames were selected as previously explained, and scene analysis was applied. In the process of scene analysis, at least three field marks for each key frame. This reference information was subsequently used as known points in order to solve the three unknown parameters that determine camera status. Note that this entire step could be avoided if a priori knowledge of the camera status was available. It is likely that in early, television network, applications of the MPI video system in coverage of structured events like American football that the camera (i) positions and (ii) status parameters will be known, and continuously known, to the MPI video system. To such extent as they are known they obviously need not be calculated.

In application of the scene analysis process to the actual video data it was found that not all video frames have enough known points. An example of a video frames that lacks sufficient known points is shown in Figure 9b. This may be contrasted with a video frame having more than sufficient known points as is shown in Figure 9a. In the experimental data used, 14 out of 15 key frames from camera 1 had at least three (3) known points, while none of seven (7) key frames from camera 2, and eight (8) out of fourteen (14) key frames from camera 3, had three (3) or more obvious known points. The difference between the cameras was that camera 1 was placed at high position while cameras 2 and 3 were placed at low positions. Accordingly, estimates had to be made for those video frames that didn't show enough obvious known points. The results of such estimations ave not necessarily accurate. Many known points an this image can be used for camera calibration.

Some examples of actual results obtained by use of the rudimentary, prototype, MPI system are shown in Figure 10.

These illustrated results were obtained by selecting "Washington" as a player to be focused on. For each video frame, a three-dimensional cursor was overlaid according to the position of "Washington". Regarding these video frames, we see that the results of scene analysis are substantially accurate according to the following observation.

First, the positions of the player "Washington" that a human may read from the video frames are close to the values that the system calculates. The values calculated by the MPI video system are shown below each picture in Fig. 10.

Second, each axis of three-dimensional cursers appears to agree with direction of the football field that a human may read from video frames.

Third, the three-dimensional cursor appear to be close to the chosen player "Washington" in the screen video image.

Other frames were checked as well. It has been confirmed that the results of the MPI video system to isolate, and to track, "target" objects of interest are mostly accurate, at least for those frames that contain enough known points to calibrate.

7. Global Multi-Perspective Perception In the MPI Video System

The present section 7 and following sections 8-__ expound the most conceptually and practically difficult portion of the MPI video system: its capture, organization and processing of real-world events in order that a system action -- such as, for example, an immediate selection, or synthesis, of an important video image (e.g., a football fumble, or an interception) -- may be predicated on this detection. Until this task is broken down into tractable parts in accordance with the present invention, it may seem to require a solution in the areas of machine vision and/or artificial intelligence, and to be of such awesome difficulty so as to likely be intractable, and impossible of solution with present technology. In fact, it is possible to make such significant progress on this task by use of modern technology applied in accordance with the present invention so as not only to get recognizable results, but so as to get results that are by some measure useful, and arguably even cost effective.

In accordance with the present invention of Multiple Perspective Interactive (MPI) video, an omniscient multi-perspective perception system based on multiple stationary video cameras permits comprehensive live recognition, and coverage, of objects and events in extended environment. The system of the invention maintains a realistic representation of the real-world events. A static model is built first using detailed a priori information. Subsequent dynamic modeling

involves the detection and tracking of people and objects in at least portions of the scene that are perceived (by the system, and in real time) to be the most pertinent.

The perception system, using camera hand-off, dynamically tracks objects in the scene as they move from one camera coverage zone to another. This tracking is possible due to several important aspects of the approach of the present invention, including (i) strategic placement of cameras for optimal coverage, (ii) accurate knowledge of scene-camera transformation, and (iii) the constraining of object motion to a known set of surfaces.

In this and the following sections of this specification, (i) a description of particularly the novel pattern and event recognition capability of the MPI video system of the present invention, and (ii) certain results presently obtainable with the system, are shown and discussed in the context of a practical implementation of the system on a college campus, to wit: a courtyard of the Engineering School at the University of California, San Diego. This environment is chosen in lieu of -- as a possible alternative choice -- further discussion of a football field and a football game because (i) it is desired to show more generally how (i) cameras may be strategically placed for optimal coverage, (ii) accurate knowledge facilitates scene-camera transformation, and (iii) object motion may be constrained to a known set of surfaces.

Momentarily considering only (iii) object motion, the exemplary courtyard environment contains (i) one object -- a human walker -- that follows a proscribed and predetermined dynamic path, namely a walkway path. The exemplary environment contains (ii) still other objects -- other human walkers -- that do not even know that they are in any of a scene, a system, or an experiment, and who accordingly move as they please in un-predetermined patterns (which are nonetheless earthbound). Finally, the exemplary environment contains (iii) an object -- a robot -- that is not independent, but which rather moves in the scene in response to static and dynamic objects and events therein, such as to, for example, traverse the scene without running into a static bench or a dynamic human.

It will therefore be recognized that even more is transpiring in the exemplary courtyard environment than on the previously-discussed football field, and that while this exemplary courtyard environment is admittedly arbitrary, it is also very rich in static and dynamic objects important to the exercise and demonstration of an omniscient multi-perspective perception capability of the MPI video system of the present invention.

7.1 Organization of the Teaching of Global Multi-Perspective Perception In the MPI Video System

Global Multi-Perspective Perception is taught and exercised in a campus environment containing a (i) mobile robot, (ii) stationary obstacles, and (iii) people and vehicles moving about -- actors in the scene that are shown diagrammatically in Figure 11a. In the present approach an omniscient multi-perspective perception system uses multiple stationary cameras which provide comprehensive coverage of an extended environment. The use of fixed global cameras simplifies visual progressing.

All dynamic objects in the environment, including the robot, can be easily and accurately detected by (i) integrating motion information from the different cameras covering these objects, and, importantly to the invention, (ii) constraining the environment by analyzing only such motion as is constrained to be to a small set of known surfaces.

The particular global multi-perspective perception system that monitors the campus environment containing people, vehicles and the robot uses the several color and monochrome CCD cameras also diagrammatically represented in Figure 11. This particular perception system is not only useful in the MPI video system, but is also useful in any completely autonomous system with or without a human in the loop, such as in the monitoring of planes on airport runways.

The operation of the global multi-perspective perception system is discussed in both human-controlled and autonomous modes. In the preferred system, individual video streams are (i) processed on separate work stations on the local network and (ii) integrated on a special purpose graphics machine on the same network. The particular system, the particular experimental setup, and pertinent performance issues, are described as follows:

The next section 8 describes the preferred approach and the principle behind camera coverage, integration and camera hand-off. The prototype global multi-perspective perception system, and the results of experiments thereon, is next described in section 9. The approach of present invention is, to the best present knowledge of inventors, a revolutionary application of computer vision that is immediately practically useable in several diverse fields such as intelligent vehicles as well as the interactive video applications -- such as situation monitoring and tour guides, etc. -- that are the principal subject of the present specification.

The applicability of the prototype global multi-perspective perception system to just some of these applications is presented in section 10. Opportunities for further improvements and expansions are discussed in Section 11.

SUBSTITUTE SHEET (RULE 26)

8. Multi-Perspective Perception

Multi-perspective perception involves each of the following.

First, the "expectations" that various objects will be observed must be generated from multiple different camera views by use of each of (i) a priori information, (ii) an environment model, and (iii) the information requirements of the present task. The statement of the immediately preceding sentence must be read carefully because the sentence contains a great deal of information, and important characterization of one aspect of the present invention. Each of (i) a priori information, (ii) an environment model, and (iii) the information requirements of the task, have variously been considered, and melded into, prior art systems for, and methods of, machine perception. Note however, that the first sentence of this paragraph is definitive. Next, note that the use of the (i) information, (ii) environment model, and (iii) information requirements is to generate -- specifically from multiple different camera views -- something called "expectations". These "expectations" are the probabilities that a (i) particular object will be observed (ii) at a particular place.

Second, objects from each camera must be independently detected and localized. This is not always done on the prior art, although it is not unduly complex. Simple motion detection is mostly used in the preferred embodiment of the present, prototype, global multi-perspective perception system.

Next, the separate observations are assimilated into a three dimensional model. In this step, the preferred embodiment of the present invention leaves "familiar ground" quickly, and "plunges" into a new construct for any perception system, whether global and/or multi-perspective or not.

Fourth, and finally, the model is used in performing the required tasks. Exactly what this means must be postponed until the "model" is better understood.

A high-level schematic diagram of the different components of the preferred embodiment of the prototype multi-perspective perception system in accordance with the present invention is shown in Figure 12. A study of the diagram will show that the system includes both two-dimensional and three-dimensional processing. Reference S. Chatterjee, R. Jain, A. Katkere, P. Kelly, D. Y. Kuramura, and S. Moezzi; Modeling and interactivity in MPI-Video, Technical Report VCL-94-103, Visual Computing Laboratory, University of California, San Diego, Dec. 1994.

Two key aspects of the architecture diagrammed in Figure 12 are the (i) static model and the (ii) dynamic model. The static model contains a priori information such as camera calibration

parameters, look-up tables and obstacle information. The dynamic model contains task specific information like two dimensional and three dimensional maps, dynamic objects, states of objects in the scene (e.g, a particular human is mobile, or the robot vehicle immobile), etc.

8.1 Three-dimensional Modeling

The three-dimensional model of the preferred embodiment of the prototype multi-perspective perception system in accordance with the present invention is created using information from multiple video streams. This model provides information that cannot be derived from a single camera view due to occlusion, size of the objects, etc. Reference S. Chatterjee, et al. op. cit.

A good three dimensional model is required to recognize complex static and moving obstacles. At a basic level, the multi-perspective perception system must maintain information about the positions of all the significant static obstacles and dynamic objects in the environment. In addition, the system must extract information from both the two-dimensional static model as well as the three-dimensional dynamic model. As such, a representation must be chosen that (i) facilitates maintenance of object positional information as well as (ii) supporting more sophisticated questions about object behavior.

While information representation can be considered an implementation issue, the particular presentation chosen will significantly affect the system development. Thus, information representation is considered to be an important element of the preferred multi-perspective perception system, and of its architecture. In the preferred system, geometric information is represented as a combination of voxel representation, gridmap representation and object-location representation. Specific implementations and domains deal with this differently.

When combined with information about the exact position and orientation of a camera, the a priori knowledge of the static environment is very rich source of information which has not previously received much attention. For each single view, the preferred system is able to compute the three dimensional position of each dynamic object detected by its motion segmentation component. To do so, the (i) a priori information about the scene and (ii) the camera calibration parameters are coupled with (iii) the assumption that all dynamic objects move on the ground surface.

Using this information it is a straightforward exercise for a practitioner of the computer programming arts to compute the equation of the line that passes through the camera projection point and a given feature on its image plane. Then, by assuming

that the lowest image point of a dynamic object is on the ground, the approximate position of the object on the ground plane os readily found. Positional information obtained from all views is assimilated and stored in the 2D grid representing the viewing area.

For the case where an object is observed by more than one camera, the three-dimensional voxel representation is particularly efficacious. Here a dynamic object recorded on an image plane projects into some set of voxels. Multiple views of an object will produce multiple projections, one for each camera. The intersection of all such projections provides an estimate of the 3-dimensional form of the dynamic object as illustrated in Figure 13 for an object seen by four cameras.

This section and its accompanying illustrations -- short as they may be -- have set forth a complete disclosure of how to make two- and three-dimensional models of the scene. It no remains only to use such models, in conjunction with other information, for useful purposes.

8.2 Automatic Camera Handoff

Camera handoff should be understood to be the event in which a dynamic object passes from one camera coverage zone to another. The multi-perspective perception system must maintain a consistent representation of an object's identity and behavior during camera handoff. This requires the maintenance of information about the object's position, its motion, etc.

Camera Handoff is a crucial aspect of processing in the multi-perspective perception system because it integrates a variety of key system components. Firstly, it relies on accurate camera calibration information, static model data. Secondly, it requires knowledge of objects and their motion through the environment determined from the dynamic model. Finally, the camera handoff can influence dynamic object detection processing. This section 8 has described the architecture, and some important features, of the multi-perspective perception system. Reference also S. Chatterjee, et al. op. cit. The next section describes in detail the preferred implementation of the multi-perspective perception system for the application of monitoring a college courtyard.

9. Setup of the Multi-perspective Perception System, and Results of System Use

The implementation of an integrated Multiple Perspective Interactive (MPI) video system demands a robust and capable implementation of the multi-perspective perception subsystem. To simplify the teaching of the multi-perspective perception

subsystem, and since this subsystem taken alone is useful in several other applications (described in Section 4) than just MPI video, the following describes the multi-perspective perception subsystem as a stand-alone system independent of the MPI video system of which it is a part. It will be understood that, one the object identifications, object tracking, and multiple perspective views of the multi-perspective perception subsystem are obtained, it is a straightforward matter to use these results in a MPI video system. (For many purposes of supplying information to the video viewer, only a high-level viewer interface is required to access the considerable current information of the multi-perspective perception subsystem.) The following sections describe the multi-perspective perception subsystem/system in detail.

9.1 Multi-Perspective Perception System Prototype

9.1.1 Setup and Use

The initial development and exercise of the multi-perspective perception system took place in a laboratory on an extended digitized color sequence. A one minute long scene was digitized from four color CCD cameras overlooking a typical campus scene 1. The one minute scene covers two pedestrians, two cyclists, and a robot vehicle moving between coverage zones. A schematic of this scene shown in Figure 14, consisting of Figure 14a and Figure 14b

For calibration and experimental evaluation of the prototype system, one of the two pedestrians walked on a pre-determined known path. No restrictions were placed on other moving objects in the scene.

9.1.2 Digitalization

The four views of the scene were digitized using a frame-addressable VCR, frame capture board combination. The synchronization was done by hand using synthetic synchronization points in the scene (known as hat drops). The resulting image sequences were placed on separate disks and controllers for independent distributed access. Having an extended pre-digitized sequence (i) accorded repeatability and (ii) permitted development of the perception system without the distractions and time consumption of repeated digitalization of the scene. The source of the scene image sequence was transparent to the perception system, and was, in fact, hidden behind a virtual frame grabber. Hence, the test was not only realistic, but migration of the perception system into (i) real-time using (ii) real video frame capture boards proved easy.

9.1.3 Camera Calibration

Calibration of the cameras in the perception system is important because accurate camera-world transformation is vital to correct system function. The cameras are assumed to be calibrated a priori, so that precise information about each camera's position and orientation could be used either directly, or by use of pre-computed camera coverage tables, to convert two dimensional observations into three dimensional model space, and, further, three dimensional expectations into 2D.

For the experimental exercise of the perception system, a complete, geometric three dimensional model of the courtyard was built using map data. This information was then used for external calibration of each camera. Calibration was done with a user in the loop. The static model was visualized from a location near the actual camera location and the user interactively modified the camera parameters until the visualized view exactly matched the actual camera view (displayed underneath).

9.1.4 Distributed Architecture

At the University of California, San Diego, cameras are physically distributed throughout the campus to provide security coverage. Because the experimental use of the perception system requires synchronized frames from these cameras at a very fast rate, frame capture was done close to the camera on separate computers. For modularity and real-time video processing, it is very important that the video be independently processed close to the sources thereof. The preferred hardware setup for the experimental exercise is pictorially diagrammed in Figure 15. Several independent heterogeneous computers -- a Sun SPARCstation models 10 and 20 and/or SGI models Indigo2, Indy and Challenge -- were selectively used based on criteria including (i) the load on the CPU, and the computer throughput, (ii) computer proximity to the camera and availability of a frame capture board (for real-time setup), and (iii) the proximity of each computer to a storage location, measured in Mbps (for the experimental setup).

The work stations in the experiment were connected on a 120 Mbps ethernet switch which guaranteed full-speed point-to-point connection. A central graphical work station was used to control the four video processing workstations, to maintain the environment model (and associated temporal database), and, optionally, to communicate results to another computer process such as that exercising and performing an MPI video function.

The central master computer and the remote slave computers communicate at a high symbolic level; minimal image information is exchanged. Hence only a very low network bandwidth is

required for master-slave communication. The master-slave information exchange protocol is preferably as follows:

First, the master computer initializes graphics, the database and the environment model, and waits on a pre-specified port.

Second, and based on the master computer's knowledge of the network, machine throughput etc., a separate computer process starts the slave computer processes on selected remote machines.

Third, each slave computer contacts the master computer, using a pre-specified machine-port combination, and an initialization hand-shaking protocol ensues.

Fourth, the master computer acknowledges each slave computer and sends the slave computer initialization information such as (i) where the images are actually stored (for the laboratory case), (ii) the starting frame and frame interval, and (iii) camera-specific image-processing information like thresholds, masks etc.

Fifth, the slave initializes itself based on the information sent by the master computer

Sixth, once the initialization is completed, the master computer, either synchronously or asynchronously depending on application, will processes the individual cameras as described in following steps seven through nine.

Seventh, whenever a frame from a specific camera needs to be processed then the master computer sends a request to that particular slave computer with information about processing the frame focus of attention windows, frame specific thresholds and other parameters, current and expected locations and identifications of moving objects etc., continuing during this processing any user interaction. In synchronous mode, requests to all slave computers are sent simultaneously and the integration is done after all slave computers have responded. In asynchronous mode, this will not necessarily proceed in unison.

Eighth, when a reply is received, the frame information is used to update the environment model and the database as described in following Section 9.1.7.

The next sections describe the communication traffic between the master and the slave computers.

9.1.5 Modeling and Visualization

A communication master computer that manages all slave computers, assimilates the processed information into an environment model, process user input (if any), and sends information to the MPI video process (if any), resides at the heart of the multi-perspective perception system. In the preferred prototype system, this master computer is an SGI

Indigo2 work station with high-end graphics hardware. This machine, along with graphics software -- OpenGL and Inventor -- was used to develop a functional Environment Model building and visualization system. Reference J. Neidev, T. Davis, and M. Woo; *OpenGL™ Programming Guide: Official Guide to Learning OpenGL, Release 1*, Addison-Wesley Publishing Company, 1993. Reference also J. Wernecke; *The Inventor Mentor: Programming Object-Oriented 3D Graphics with Open Inventor™; Release 2*, Addison-Wesley Publishing Company, 1994.

In the preferred system, Inventor manages the scene database and OpenGL performs the actual rendering. A "snapshot" view of the visualization system of the master computer, including four camera views, and a rendered model showing all the moving objects in iconic forms, is shown in Figure 18.

9.1.6 Video Processing

One of the goals of the exercise of the multi-perspective perception system was to illustrate the advantages of using static cameras for scene capture, and the relative simplicity of visual processing in this scenario when compared to processing from a single camera. While more sophisticated detection, recognition and tracking algorithms are still being developed and applied, the initial, prototype multi-perspective perception system uses simple yet robust motion detection and tracking. In the prototype system, and as described in previous sections, the processing of individual video streams is done using independent video processing slaves, possibly running on several different machines. The synchronization and coordination of these slaves, any required resolution of inconsistencies, and generation of expectations is done at the master.

Independent processing of information streams is an important feature of the information assimilation architecture of the present invention, and is a continuation and outgrowth of the work of some of the inventors and their colleagues. See, for example, R. Jain; *Environment models and information assimilation*, Technical Report RJ 6866(65692), IBM Almaden Research Center, San Jose, CA, 1989; Y. Voth and R. Jain; *Knowledge caching for sensor-based systems*, Artificial Intelligence, 71:257-280, Dec. 1994; and A. Katkere and R. Jain; *A framework for information assimilation*, to be published in Exploratory Vision edited by M. Landy, et al., 1994.

The independent processing results in pluggable and dynamically reconfigurable processing tracks. The preferred, prototypical, communication slave computers perform the following steps on each individual video frame. Video processing is limited by focus of attention rectangles specified by the master computer, and pre-computed static mask images

delineating portions of a camera view which cannot possibly have any interesting motion. The computation of the former is done using current locations of the object hypotheses in each view and projected locations in the next view. The latter is currently created by hand, painting out areas of each view not on the navigable surface (walls, for example). Camera coverage tables help the master computer in these computations. Coverage tables, and the concept of objects, are both illustrated in Figure 16.

In operation, the input frame is first smoothed to remove some noise. Then the difference image $d_{t,t-1}$ is computed as follows. Only pixels that are in the focus of attention windows and that are not masked are considered.

$$d_{t,t-1} = Threshold\ (\ Abs\ (\ F_{t-1} - F_t\ ),\ threshold\_value\ )$$

Optionally, to remove motion shadows, following operation is done:

$$d_t^{\pm} = d_{t,t-1} \,\&\, d_{t,t+1}$$

This shadow-removing step is not invariably used nor required since it needs a one frame look-ahead. In many cases simple heuristics may be used to eliminate motion shadows at a symbolic level.

Nest, components on binary difference image are computed based on a four-neighborhood criterion. Components that are too small or too big are thrown away because they usually constitute noise. Frames that contain a large number of components are also discarded. Both centroid (from first moments), and orientation and elongation (from the second moments), are extracted for each component.

Next, several optional filters are applied at the slave site to the list of components obtained from the previous step. Commonly used filters include (i) merging of overlapping bounding boxes, (ii) hard limits of orientation and elongation, and (iii) distance from expected features etc.

Finally, the resulting list is sent back to the master site.

9.1.7 Assimilation and Updating Object Hypotheses

The central visualization and modeling site receives processed visual information from the video processing sites and creates/updates object hypotheses. There are several

sophisticated ways of so doing. Currently, and for the sake of simplicity in developing a completely operative prototype, this is done as follows:

First, he list of two-dimensional (2-D) object bounding boxes is further filtered based on global knowledge.

Second, the footprint of each bounding box is projected to the primary surface of motion by intersecting a ray drawn from the optic center of that particular camera through the foot of the bounding box with the ground surface.

Third, each valid footprint is tested for membership with existing objects and the observation is added as support to the closest object, if any. If no object is close enough, then a new object hypothesis is created.

Fourth, all supporting observations are used (with appropriate weighting based on distance from the camera, direction of motion, etc.) to update the position of each object.

Fifth, the object positions are projected into the next frame based on a domain-dependent tracker.

Sixth, if events in the scene are to be recognized, object positions and associations are compared against predetermined templates. For example, if in the courtyard scene the robot has moved into spatial coincidence with one of the predetermined immovable objects, such as a bench, then the robot may have run into the bench -- an abnormal and undesired occurrence. For example, if in the scene of a football game the football has moved in a short time interval from spatial coincidence with a moving player that was predetermined to be of a first team to spatial coincidence with a moving player that is predetermined to be of a second team -- especially if the football is detected to have reversed its direction of movement on the field -- then any of a (i) kickoff, (ii) fumble, or (iii) interception may have transpired. If the detected event is of interest to the viewer in the MPI video system, then appropriate control signals are sent. Also, based on the sub-systems knowledge of static objects, if an actual or projected position of a dynamic object intersects a static object, then an appropriate message may be sent. If the scene of a football game the football is determined to be in spatial coincidence with the forty yard marker, then it is reported that the football is on the forty yard line.

9.1.8 Results

Each of Figures 17 through 21 frames in an exemplary exercise -- consisting of one thousand (1000) total frames from four (4) different cameras acquired as described in Section 9.1.2 -- of the Multi-perspective perception subsystem.

Figures 17 through 19 show the state of the subsystem at global time 00:22:29:06. Figures 20 and 21 show the state of the subsystem at the global time 00:22:39:06. In Figure 17, four dynamic objects are shown in the scene: a robot vehicle, two pedestrians and a bicyclist. The scene is covered by four different cameras. A fifth object -- another bicyclist -- is shown, but is not labeled for clarity.

Each of the four cameras has its own clock, as is shown under the camera's view in one of Figures 17 through 17d. Camera number three (#3), which is arbitrarily known as "Saied's camera", is used to maintain the global clock since this camera has the largest coverage and the best image quality. Figure 17a-17d clearly shows the coverage of each camera.

As shown in Figure 17, an object that is out of view, too small, and/or occluded from view in one camera is in view, large and/or un-occluded to the view of another camera. Note that the object labels used in the Figure 17 are for explanation only. The prototype subsystem does not include any non-trivial object recognition, and all object identifiers that persist over time are automatically assigned by the system. Mnemonic names like "Walker 1", or "Walker" refer to the same object identification (e.g., what the software program would label "BasicEnvObject0023", "BasicEnvObject0047", etc.) over all the different frames of Figures 17-21.

A pictorial representation of the display screen showing the operator interface to the multi-perspective perception subsystem is shown in Figure 18. Four camera views are shown in the top row of Figure 18. Each view is labeled using its mnemonic identification instead of its numeric identification because humans respond better to mnemonic "id's". Each view may be associated with a one of Figures 17a-17d.

A red rectangle is drawn automatically around each detected object in each camera's view of the scene. It can be clearly seen how objects are robustly detected in the different images obtained with cameras of different characteristics (huge variations in color, color vs. monochrome) -- even when the object is just a few pixels wide.

The bottom section of the operator display screen in Figure 18 shows the object hypotheses which are formed over several frames (first frame is global clock 00:22:10:0). The intensity each object's marker represents the confidence in each hypotheses. The entire display screen, the objects depicted, and the object hypothesis diagrammatically depicted, is, as might well be expected, in full color. Figures 17-21 are therefore monochrome of color images. In particular, the object markers are preferably in the color yellow, and the intensity of the bight yellow color of each object's marker represents the

confidence in the hypotheses for that object. The eye is sensitive to discern even such slight differences in color intensity as correspond to differences in confidence.

The multi-perspective perception subsystem has a high confidence in each object for which a marker is depicted in Figure 18 because, at the particular global time represented, each object happens to have been observed from many cameras over several past frames.

The three-dimensional model at global time 09:22:29:06 is shown in Figures 19a-19e in both real and virtual views. Figures 19a-19d show the model from the four real camera views. One-to-one correspondence between the model and the camera views can be clearly seen. The fifth view of Figure 19e is a virtual view of the model from directly overhead the courtyard -- where no real camera actually exists. This virtual view shows the exact locations of all three objects, including the robotic vehicle, in the two-dimensional plane of the courtyard. Three objects are very accurately localized. The fourth object, Walker Number Two (#2) in Figure 17 and 19, has some error in localization since this person is (i) not visible in Camera number four (#4), and (ii) his/her coverage is very small in Cameras numbers two and three (#2 & #3), hence leading to some errors.

Note that even though the object Walker number two (#2) 2 is visible in Camera number one (#1), that particular observation is not used since its bounding box intersects the bottom of the image. Obviously, when an object's bounding box intersects the bottom of the image, its full extent cannot be determined and should be ignored. To show the development of object hypotheses over time, a snapshot of the experiment is taken ten (10) seconds later. Figures 20 and 21 show that state. Figure 20 corresponds to Figure 18 while Figure 21 corresponds to Figure 19. One important observation to make in Figures 20 and 21 is that, given the relative proximity of Walker number one (#1) and Bicyclist number one (#1), both are still classified as separate objects. This is only possible due to the subsystem's history and tracking mechanism.

9.2 Applications

In addition to multi-perspective interactive (MPI) video, a variety of other application areas can benefit from the global multi-perspective perception subsystem described. For instance, environments demanding sophisticated visual monitoring, such as airport runways and hazardous or complex roadway traffic situations can advantageously use the global multi-perspective perception subsystem. In these environments, as in MPI video, objects must be recognized and identified, and spatial-temporal

information about objects' locations and behaviors must provided to a user.

The expected first application of the global multi-perspective perception subsystem to the MPI video system has been in sports, and it is expected that sports and other entertainment applications -- which greatly benefit -- will be the first commercial application of the subsystem/system. Sports events, e.g. football games, are already commonly imaged with video cameras from several different spatial perspectives -- as many as several dozen such for a major professional football game. The reason that still more cameras are not used is primarily perceived as having to do to the expense of such human cameramen as are required to focus the camera image on the "action", and not the cost of the camera. Additionally, it is unsure how many different "feeds" a sports editor can use and select amongst -- especially in real time. The reason the televised sporting event viewing public is by an large satisfied with the coverage offered is that they have never seen anything better -- including in the movies. Few people have been privileged to edit a movie or a video, and even fewer to their own personal taste (no matter how weird, or deviant). The machine-based MPI video of the present invention will, of course, accord viewing diversity without the substantial expense of human labor.

Consider that, in using the global multi-perspective perception subsystem and the MPI video system, multiple video perspectives are integrated into a single comprehensive model of the action. Such a representation can initially assist a number of video editors in choosing between different perspectives, for example a video editor for the "defense", and one for the "offense" and one for the "offensive receivers", etc., as well as the standard "whole game" video editor. Ultimately, and with increasingly affordable computer power, even a regular viewer who is interested, for example, in a particular player would be able to customize his video display based on that player. Interactive Video applications such as these will greatly benefit from, and will use, both the global multi-perspective perception subsystem and the MPI video system.

Still another application where the global multi-perspective perception subsystem may be used directly is as a tour guide in a museum or any such confined space. Rather than moving objects in the scene (i.e, the courtyard, or the football field), the scene can remained fixed (i.e., the museum) and the camera can move. The response accorded a museum visitor/video camera user will be even more powerful than, for example, the hypertext linkage on the World Wide Web of the Internet. On an interactive computer screen and system (whether

on the Internet or not) a viewer/user and point and click his/her way to additional information. However, the viewer/user is viewing on a video representation of museum art, and not the real thing.

Consider now a visit to a museum of art using, instead of a self-guided tour headset, a hand-held video camera. The user/viewer can go anywhere that he or she wants within the galleries of the museum, and can point at any art work, to perhaps show not only the scene at hand in the viewfinder of his or her video camera, but perhaps also a video and/or audio overlay that has interactively been sent to the user's video camera from "computer central". The "computer central" recognizes where in the museum the user's video -- which is also transmitted out to the "computer central" -- arises from. Simple "helps" in the gallery rooms, such as bar codes, may perhaps help the "computer central" to better recognize where an individual user is, and in what direction the user is pointing. So far this scheme may not seem much different, and potentially more complex and expensive, than simply having a user-initiated information playback system at each painting (although problems of time synchronization for multiple simultaneous viewers may be encountered with such a system).

The advantage that the global multi-perspective perception subsystem offers in the art museum environment is that accumulation of a "user track", instead of an "object track", becomes trivial. The user may be guided in a generally non-repetitious track through the galleries. If he/she stops and lingers for a one artist, or a one subject matter, or a style, or a period, etc., then selected further works of the artist, subject matter, style, period, etc., that seem to command the user's interest may be highlighted to the user. If the user dwells at length at a single work, or at a portion thereof, the central computer can perhaps send textual or audio information so regarding. If the user fidgets, or moves on, then the provided information is obviously of no interest to the user, and may be terminated. If the user listens and views through all offered messages that are classified "historical perspective of the persons and things depicted in the art work viewed", then it might reasonably be assumed that the user is interested in history. If, on the contrary, the user listens and views through all offered messages that are classified "life of the artist", then it might reasonably be assumed that the user is interested in biography.

9.3 Conclusions, and Future Developments, Concerning the Global Multi-perspective Perception Subsystem

The complex phenomena of "man-machine information systems

of the future" discussed in the immediately proceeding section may seem all "fine and good", or even fascinating, but some minutes deliberation are likely required to understand exactly what this all has to do with the present invention. In the simplest possible terms, information -- and a great, great deal of such information, indeed -- comes to a camera, which is the best present machine substitute for human vision, in the form of two-dimensional images. However, our own human vision is stereoscopic, and our eye/brains combination, perceptive of not two, but three, dimensions. We reason things out spatially in three dimensions, and we are interested in what goes on in three dimensions -- as at a real live football game -- as well as in two dimensions -- as in the presentation of a football game on television. (We are also interested in smelling, tasting and/or hearing concurrently with our viewing, but the present invention cannot do anything about satisfying this desire.)

It is the teaching of the present invention, broadly speaking, that in order to best serve man, machine systems that convey visual information ought to, if at all possible or practical, "rise to the level" of three-dimensional information. The machine system would desirably so rise not in the images that it displays to viewers (which displayed images will, alas, remain two-dimensional for the foreseeable future) but, instead, in the construction and management of a database from which information can be drawn. Moreover, if this three-dimensional database is good enough, and if the machine (computer) processes that operate upon it are clever enough, then the power, and the flexibility, or viewer service, and presentations, are magnified. This magnification is in the same sense that we get more out of life by operating as autonomous agents in the three-dimensional world than we would if we could view all the cinema of the world for free forever in a darkened room. If a human cannot interact with his/her environment -- even as viewed, when necessary, through a two-dimensional window -- then some of the essence of living is surely lost.

It is the teaching of the present invention how to so construct from multiple two-dimensional video images a three-dimensional database, and how to so manage the three-dimensional database for the production of two-dimensional video images that not necessarily those images from which the database was constructed.

Future improvements to the global multi-perspective perception subsystem will involve building on the complete framework provided in this specification. Improvements on two dimensional motion detection and tracking, three dimensional integration and tracking, etc. are possible. Another important extension of the present invention would be to use cooperative

active cameras for enhanced track robots and other moving objects over wide areas. This approach could both (i) reduce the number of cameras required to cover an area, and (ii) improve object detection and recognition by keeping objects towards center of view.

Future improvements to the global multi-perspective perception subsystem may also be taken in the area of cooperative human-machine systems. Interactivity at the central site might be improved so as to permit a human to perform higher-level cognitive tasks than simply asking "where", or "what/who?", or "when". The human might ask, for example, "why?". In the context of football, and for the event of a tackle, the machine (the computer) might be able to advance as a possible answer (which would not invariably be correct) to the question "why (the tackle)?" something like: "Defensive Linebacker #24 at the (site of) tackle has not been impeded in his motion since the start of the play.". The machine has sensed that linebacker #24 -- who may or may not have actually made the tackle but who was apparently nearby -- was not in contact with any defensive player prior to the tackle. In a highest-level interpretation of this event as would be, and as of the present can be, rendered only by a human being, the likely interpretation of this sequence -- as was recognized by the machine -- is that someone has missed a tackle.

10. <u>The Particular, Rudimentary, Embodiment of the Invention Taught Within This Specification</u>

The present specification has taught a coherent, logical, and useful scheme of implementing virtual video/television. The particular embodiment within which the invention is taught is, as would be expected and as is desirable for the sake of simplicity of teaching, rudimentary.

The rudimentary nature of the particular embodiment taught within this specification dictates, for example, that the described manipulation and synthesis is of recorded video images, and is not of television in real time. However, this factor is a function only of the power of the computer used. The efficacy and utility of the image manipulation and synthesis scheme of the present invention taught, including by rigorous mathematics, is not diminished by the computational speed at which it is accomplished.

The rudimentary nature of the particular embodiment taught within this specification further dictates, for example, that the extraction of some scene features from these video images is not only not in real time, but is in fact done manually. This will turn out to be an insignificant expedient. First, many of the features extracted will turn out to be (i) distinct and (ii)

fixed; and are in fact the hash marks and yard markings of an American football field! It is clear that these fixed features could be entered into any system, even by manual means, just once "before the game". Moreover, they are easily captured by even the most rudimentary machine vision programs. Other features extracted from the video images -- such as football players and/or a football in motion -- are much harder to extract, especially at high speeds and most especially in real time. To extract these moving features enters the realm of machine vision. Nonetheless that this portion of the system of the present invention is challenging, many simple machine solutions -- ranging from fluorescently bar-coded objects in the scene (e.g., players and football) to full-blown, state-of-the-art machine vision programs -- are possible and are discussed within this specification. In fact, with non-real-time video it is even possible -- and quite practical -- to have a trained human, or a squad of such, track each player or other object of concern through each video scene (e.g., a football play). The "tracked" objects (the players) are only viewed later, upon an "instant replay" or from a video archive on tape or CD-ROM. Accordingly, it is respectfully suggested that the utility, and the scope, of the present invention is not degraded by certain practical limitations, as of present, on the particular image extraction function performed in the rudimentary embodiment of the invention.

Finally, in the particular, rudimentary, embodiment of the invention taught in this specification the synthesized video image is not completely of a virtual camera/image that may be located anywhere, but is instead of a machine-determined most appropriate real-world camera. This may initially seem like a significant, and substantive, curtailment of the described scope of the present invention. However, important mitigating factors should be recognized. First, the combination of multiple images, even video images, to generate a new image is called "morphing", and is, circa 1995, well known. One simple reason that the rudimentary system of the present invention does proceed to perform this "well known" step is that it is slow when performed on the engineering workstation on which the rudimentary embodiment of the present invention has been fully operationally implemented. Another simple reason that the rudimentary system of the present invention does proceed to perform this "well known" step is that, for the example of American football initially dealt with by the system and method of the present invention, it is uncertain whether this expensive, and computationally extensive, step (which turns out to be a final step) is actually needed. Namely, many cameras exist, and will exist, at a football telecast. Even if some

virtual image is desired of, for example, the right halfback during the entirety of one play, it is likely that some existing camera or combinations thereof can deliver the desired image(s). Accordingly, it is again respectfully suggested that the utility, and the scope, of the present invention is not degraded by certain practical limitations, as of present, on the particular selection/morphing function performed in the rudimentary embodiment of the invention.

In return for some compromises rooted in practical considerations, the present specification completely teaches, replete with pictures, how to implement a virtual video camera, and a virtual video image, by synthesis in a computer and in a computer system from multiple real video images that are obtained by multiple real video cameras. Because this synthesis is computationally intensive, the computer is usefully powerful, and is, in the preferred embodiment, an engineering workstation.

Moreover, depending upon how extensively and how fast (i) three-dimensional analysis of the multiple scenes is to transpire, (ii) information from the multiple scenes is to be extracted, and (iii) linkage between the multiple scenes is to be established, the computer and computer system realizing the present can usefully be very powerful, and can usefully exercise certain exotic software functions in the areas of machine vision, scene and feature analysis, and interactive control.

As explained, the present invention has not been, to the present date of filing, implemented at its "full blown" level of interactive virtual television. It need not be in order that it may be understood as a coherent, logical, and useful scheme of so implementing virtual video/television.

10.1 Directions of Future Development

This specification has described the development and actual use of a prototype football video retrieval system. This system serves to demonstrate the concepts and the potential of MPI video. The feasibility of the broader concepts is completely demonstrated. Design and implementation of MPI video for longer sequences of football, and also for other applications, is still proceeding as of the filing date.

However, as is also clear from the present specification, the MPI video system is in its infancy. The potential of the MPI video techniques is obvious, but cost effective implementation, especially for the individual "John Q. Public" viewer has a long way to go. Almost all medium- to large-scale
computer technology involved in the implementation of the prototype MPI video systems was stretched to its limits. The following are only a few examples of the useful, and probable,

future developments and enhancements.

10.1.1 Scene Analysis

In the prototype MPI video system, much information was inserted manually by an operator. However to make MPI video practical for commercial use, this process should be automated as much as possible. (Notice that it is not necessary that MPI video should invariably be so automated in order to be used. Certain very crucial or interesting events for which multiple video images exist -- such as key plays in sporting events -- may be well deserving of careful analysis after the fact.)

Also, and as may be recalled, it was found to be difficult to determine camera status for some video frames which contain very few known points to calibrate. This problem may be solved by using information obtained from other video frames, both of other cameras in the same instant and/or of the same camera in the instants before and after. Once this technology becomes practical, it will be possible to structure many other items and objects to simplify the object recognition task.

10.1.2 Data Modeling and Indexing

Information structure that is contained in a scene is usually complicated, and the amount of information in the scene is huge. Moreover, this video information is developed and received over but a short period of time. To deal with various types of queries, good data modeling is required. See Amarnath Gupta, Terry Weymouth, and Ramesh Jain; "Semantic queries with pictures: the VIMSYS model" appearing in *Proceedings of the 17th International Conference on Very Large Data Bases*, September 1991.

To enable the best quick response to the queries, indexing techniques will be required. These techniques for images and video are just being developed.

10.1.4 The Human Interface

The present specification has taught that interaction using three-dimensional cursor is a good way for a user/viewer to point or highlight objects in three-dimensional space. However, in the field of entertainment and training, where interactive video is expected to be useful, an even more friendly interface is desired. Techniques to specify camera location, describe events of interest, and other similar things need further development. In many applications, like "telepresence", one may require extensive use of virtual reality environments. In applications like digital libraries, strong emphasis on user modeling will be essential.

Nonetheless to the potential of improving, and rendering

more abstract, the user/viewer interface in some applications, this interface is most assuredly not a "weak point" of the present invention of MPI video. Indeed, it is difficult to even imagine how new and improved user/viewer interface tools may be used in the context of interactive movies and similar other applications of MPI video. It seems as if the tools that the user/viewer might reasonably require are already available right now.

10.1.4 Video Databases

As access to data from more and more cameras is permitted, the storage requirements for MPI video will increase significantly. Where and how to store this video data, and how to organize it for timely retrieval, is likely to be a major issue for expansion and extension of the MPI video system. In the prototype system, the single most critical problem has been the storage of data. Future MPI video will continue to put tremendous demands on the capacity and efficiency of organization of the storage and database systems.

10.2 Recapitulation of the MPI Portion of the Present Invention

In one, rudimentary, embodiment of present invention, a virtual video camera, and a virtual video image, of a scene were synthesized in a computer and in a computer system from multiple real video images of the scene that were obtained by multiple real video cameras.

This synthesis of a virtual video image was computationally intensive. Depending upon how extensively and how fast (i) three-dimensional analysis of the multiple scenes is to transpire, (ii) information from the multiple scenes is to be extracted, and (iii) linkage between the multiple scenes is to be established, the computer and computer system realizing the present can usefully be very powerful, and can usefully exercise certain exotic software functions in the areas of machine vision, scene and feature analysis, and interactive control. In the prototype system network-connected engineering work stations that were relatively new as of the 1995 date of filing were used.

Notably, however, the present invention need not be (and to the present date of filing has not been) implemented at its "full blown" level of interactive virtual television in order that it may be recognized that a coherent, logical, and useful scheme of implementing virtual video/television is shown taught.

The virtual video camera, and virtual image, produced by the MPI video system need not, and commonly does not, have any real-world counterpart. The virtual video camera and virtual

image may show, for example, a view of a sporting event, for example American football, from an aerial, or an on-field, perspective at which no real camera exists or can exist.

In advanced, computationally intensive, from the virtual camera/virtual image can be computer synthesized in real time, producing virtual television.

The synthesis of virtual video images/virtual television pictures may be linked to any of (i) a perspective, (ii) an object in the video/television scene, or (iii) an event in the video/television scene. The linkage may be to a static, or a dynamic, (i) perspective, (ii) object or (iii) event. For example, the virtual video/television camera could be located (i) statically at the line of scrimmage, (ii) dynamically behind the halfback wheresoever he might go, or (iii) dynamically on the football wheresoever it might go, in a video/television presentation of a game of American football.

The virtual camera, and virtual image, that is synthesized from multiple real world video images may be so synthesized interactively, and on demand. For example, and in early deployments of the system of the invention, a television sports director might select a virtual video replay of a play in a football game keyed on a perspective, player or event, or might even so key a selected perspective of an upcoming play to be synthesized in real time, and shown as virtual television. Ultimately, many separate viewers are able to select, as sports fans, their desired virtual images. For example, a virtual video replay, or even a virtual television, image of each of the eleven players on each of two American football teams, plus the image of the football, is carried on twenty-three television channels. The "fan" can thus follow his favorite player.

Ultimate interactive control where each "fan" can be his own sports director is possible, but demands that considerable image data (actually, three-dimensional image data) be delivered to the "fan" either non-real time in batch (e.g., on CD-ROM), or in real time (e.g., by fiber optics), and, also, that the "fan" should have a powerful computer (e.g., an engineering workstation, circa 1995).

In accordance with the preceding explanation, variations and adaptations of Multiple Perspective Interactive (MPI) video in accordance with the present invention will suggest themselves to a practitioner of the digital imaging arts. For example, monitors of the positions of the eyes might "feed back" into the view presented by the MPI video system in a manner more akin to "flying" in a virtual reality landscape than watching a football game -- even as a live spectator. It may be possible for a viewer to "swoop" onto the playing field, to "circle" the stadium, and even, having crossed over to the "other side" of

the stadium, to pause for a look at that side's cheerleaders.

11. Immersive Video, and the Motivation for Immersive Video

Because it provides a comprehensive visual record of environment activity, video data is an attractive source of information for the creation of "virtual worlds" which, nonetheless to being virtual, incorporate some "real world" fidelity. The present invention concerns the use of multiple streams of video data for the creation of immersive, "visual reality", environments.

The immersive video system of the present invention for so synthesizing "visual reality" from multiple streams of video data is based on, and is a continuance of the Multiple Perspective Interactive Video (MPI-Video) just discussed. An immersive video system incorporates the MPI-Video architecture, which architecture provides the infrastructure for the processing and the analysis of multiple streams of video data.

The MPI-Video portion of the immersive video system (i) performs automated analysis of the raw video and (ii) constructs a model of the environment and object activity within the environment. This model, together with the raw video data, can be used to create immersive video environments. This is the most important, and most difficult, functional portion of the immersive video system. Accordingly, this MPI-Video portion of the immersive video system is first re-visited, and actual results from an immersive "virtual" walk through as processed by the MPI-Video portion of the immersive video system are presented.

As computer applications that model and interact with the real-world increase in numbers and types, the term "virtual world" is becoming a misnomer. These applications, which require accurate and real-time modeling of actions and events in the "real world" (e.g., gravity), interact with a world model either directly (e.g., "telepresence") or in a modified form (e.g., augmented reality). A variety of mechanisms can be employed to acquire data about the "real world" which is then used to construct a model of the world for use in a "virtual" representation.

Long established as a predominant medium in entertainment and sports, video is now emerging as a medium of great utility in science and engineering as well. It thus comes as little surprise that video should find application as a "sensor" in the area of "virtual worlds." Video is especially useful in cases where such "virtual worlds" might usefully incorporate a significant "real world" component. These cases turn out to be both abundant and important; basically because we all live in, and interact with, the real world, and not inside a computer video game. Therefore, those sensations and experiences that are most valuable, entertaining and pleasing to most people most of the time are sensations and experiences of the real world, or at least sensations and

experiences that have a strong real-world component. Man cannot thrive on fantasy alone (which state is called insanity); a good measure of reality is required.

In one such use of video as a "sensor", multiple video cameras cover a dynamic, real-world, environment. These multiple video data streams are a useful source of information for building, first, accurate three-dimensional models of the events occurring in the real world, and, then, completely immersive environments. Note that the immersive environment does not, in accordance with the present invention, some straight from the real world environment. The present invention is not simply a linear, brute-force, processing of two-dimensional (video) data into a three-dimensional (video) database (and the subsequent uses thereof). Instead, in accordance with the present invention, the immersive environment comes to exist through a three-dimensional model, particularly a model of real-world dynamic events. This will later become clearer such as in, inter alia, the discussion of Figure 25.

In the immersive video system of the present invention, visual processing algorithms are used to extract information about object motion and activity (both of which are dynamic by definition) in the real world environment. This information -- along with (i) the raw video data and (ii) a priori information about the geometry of the environment -- is used to construct a coherent and complete visual representation of the environment. This representation can then be used to construct accurate immersive environments based on real world object behavior and events. Again, the rough concept, if not the particulars, is clear. The immersive environment comes to be only through a model, or representation, or the real world environment.

While video data proves powerful source medium for these tasks (leading to the model, and the immersive environment), the effective use of video requires sophisticated data management and processing capabilities. The manipulation of video data is a daunting task, as it typically entails staggering amounts of complex data. However, in restricted domains, using powerful visual analysis techniques, it is possible to accurately model the real world using video streams from multiple perspectives covering a dynamic environment. Such "real-world" models are necessary for "virtual world" development and analysis.

The MPI-Video portion of the immersive video system builds the infrastructure to capture, analyze and manage information about real-world events from multiple perspectives, and provides viewers (or persons interacting with the scene) interactive access to this

information. The MPI-Video sub-system uses a variety of visual computing operations, modeling and visualization techniques, and multimedia database methodologies to (i) synthesize and (ii) manage a rich and dynamic representation of object behavior in real-world environments monitored by multiple cameras (see Figures 2 and 22).

An Environment Model (EM) is a hierarchical representation of (i) the structure of an environment and (ii) the actions that take place in this environment. The EM is used as a bridge between the process of analyzing and monitoring the environment and those processes that present information to the viewer and support the construction of "immersive visual reality" based on the video data input.

The following sections explain the use of multiple streams of video data to construct "immersive visual reality" environments. In addition, salient details are provided regarding support of the MPI-Video subsystem for other video analysis tasks.

A variety of design issues arise in realizing immersive environments, and in managing and processing of multiple streams of video data area. These issues include, for instance, how to select a "best" view from the multiple video streams, and how to recognize the frame(s) of a scene "event". Interactively presenting the information about the world to the viewer is another important aspect of "immersive visual reality". For many applications and many viewer/users, this includes presentation of a "best" view of the real-world environment at all times to the viewer/user. Of course, the concept of what is "best" is dependent on both the viewer and the current context. In following Section 12, the different ways of defining the "best" view, and how to compute the "best" view based on viewer preferences and available model information, is described.

In some applications, e.g., "telepresence" and "telecontrol", immersion of the viewer/user is vital. Selecting the "best" view among available video streams, which selection involves constant change of viewer perspective, may be detrimental towards creating immersion. Generalizing the "best" view concept to selecting a continuous sequence of views that best suit viewer/user requirements and create immersion overcomes this. When such arbitrary views are selected, then the world must somehow be visualized from that perspective for the viewer/user.

Traditionally, immersion has been realized by rendering three-dimensional models realistically, preferably in stereo. This is the approach of the common computer game, circa 1995, offering "graphics immersion". This approach, which uses a priori texture

maps, suffers from some defects when the immersive experience to be created is that of a real-world environment. In real-world environments, the lighting conditions change constantly in ways that cannot be modeled precisely. Also, unknown dynamic objects can appear, and when they do it is not clear how and what to render.

When multiple video cameras covering an environment from multiple perspectives, as in the immersive video system of the present invention, than, in accordance with the invention, video can be used as a dynamic source of generating texture information. The complete immersive video system discussed in Section 13 uses comprehensive three dimensional model of the environment and the multiple video channels to create immersive, realistic renditions of real-world events from arbitrary perspective in both monocular and stereo presentations.

The further sections of this specification are organized as follows: Section 12 is a description of the construction of accurate three dimensional models of an environment from multi-perspective video streams in consideration of a priori knowledge of an environment. Specifically, section 12 discusses the creation of an Environment Model and also provide details on the preferred MPI-Video architecture.

Following this, section 4 describes how this model, along with the raw video data, can be used to build realistic "immersive visual reality" vistas, and how a viewer can interact with the model.

Details on the implementation of the MPI-Video portion of the immersive video system, outlining hardware details, etc., are given in section 14.

The possibilities of using video to construct immersive environments are limitless. Section 15 describes various applications of the immersive video system of the present invention.

12. Applications of Video-Based Immersive Environments

It is the contention of the inventors that video of real-world scenes will play an important role in automation and semi-automation of both (i) virtual and (ii) immersive visual reality environments. In telepresence applications, a virtual copy of the world is created at a remote site to produce immersion. See B. Chapin, *Telepresence Definitions*, a World Wide Web (WWW) document on the Internet at URL http://cdr.stanford.edu/html/telepresence/definition.html, 1995.

Key features of telepresence applications are: 1) the entire application is real-time; 2) the virtual world is reasonably faithful to the real world being mimicked; 3) since real-time and real-world are cardinal, sensors should be used in acquiring the virtual world in a completely automated way; and 4) the virtual world must be visualized realistically from the viewer perspective.

The MPI-Video modeling system described in Section 12 uses multiple video signals to faithfully reconstruct a model of the real-world actions and structure. A distributed implementation coupled with expectation-driven, need-based analysis (described in Section 14) ensures near real-time model construction. The preferred immersive video system, described in Section 13, reconstructs realistic monocular and stereo vistas from the viewer perspective (see, for example, Figure 33).

Even in non-real time applications, video-based systems, such as the one taught in this specification, can be very beneficial. Generally, it is very difficult and laborious to construct virtual environments by hand. In a semi-autonomous mode, however, a video-based system can assist the user by assuming the low level tasks like building the structural model based on the real-world, leaving only high level annotation to the user.

Video data can be used to collect a myriad of visual information about an environment. This information can be stored, analyzed and used to develop "virtual" models of the environment. These models, in turn can be analyzed to determine potential changes or modifications to an environment. For instance, MPI-Video might be employed at a particularly hazardous traffic configuration. Visual data of traffic would be recorded and analyzed to determine statistics about usage, accident characteristics, etc. Based on this analysis, changes to the environment could be designed and modeled, where input to the model again could come from the analysis performed on the real data. Similarly, architectural analysis could benefit by the consideration of current building structures using MPI-Video. This analysis could guide architects in the identification and modeling of building environments.

13. MPI-Video Architecture

To effectively create synthetic worlds which integrate real and virtual components, sophisticated data processing and data management mechanisms are required. This is especially true in the case where video is employed because high frame rates and large images result in daunting computational and storage demands. The

present invention address such data processing and management issues through the concept of Multiple Perspective Interactive Video (MPI-Video).

MPI-Video is a framework for the management and interactive access to multiple streams of video data capturing different perspectives of the same or of related events. As applied to the creation of virtual environments, MPI-Video supports the collection, processing and maintenance of multiple streams of data which are integrated to represent an environment. Such representations can be based solely on the "real" world recorded by the video cameras, or can incorporate elements of a "virtual" world as well.

The preferred MPI-Video system supports a structured approach to the construction of "virtual worlds" using video data. In this section the MPI-Video architecture, shown in Figure 1, is outlined. Those elements salient to the application of MPI-Video in the context of the processing and creation of "immersive visual reality" are highlighted.

In brief, MPI-Video architecture involves the following operations. During processing, multiple data streams are forwarded to the Video Data Analyzer. This unit evaluates each stream to (i) detect and track objects and (ii) identify events recorded in the data. Information derived in the Video Data Analyzer is sent to the Assimilator. Data from all input streams is integrated by the Assimilator and used to construct a comprehensive representation of events occurring in the scene over time (e.g. object movements and positions).

The Assimilator thus models spatial-temporal activities of objects in the environment, building a so-called environment model. In addition, these tracking and modeling processes provide input to the database which maintains both the annotated raw video data as well as information about object behavior, histories and events. Information in the database can be queried by the user or by system processes for information about the events recorded by the video streams as well as being a data repository for analysis operations. A View Selector module -- used to compute and select "best views" and further discussed below -- interfaces with the database and a user interface subsystem to select appropriate views in response to user or system input.

A visualizer and virtual view builder uses the raw video data, along with information from the environment model to construct synthetic views of the environment.

Finally, a user interface provides a variety of features to

SUBSTITUTE SHEET (RULE 26)

support access, control and navigation of the data.

To demonstrate and explore the ideas involved in MPI-Video, a prototype system was constructed. The prototype system uses data from a university courtyard environment. Figure 22a shows a schematic of this courtyard environment, indicating the positions of the cameras. Synchronized frames from each of the four cameras are shown in Figures 22b and 22c.

13.1 Three-Dimensional Environmental Model

"Virtual worlds" -- whether of an actual "real world" environment or a purely synthetic environment -- depend on the creation and maintenance of an Environment Model (EM). The EM will be understood to be a comprehensive three-dimensional model containing both (i) the structural primitives of the static environment, e.g. surfaces, shapes, elevation, and (ii) characteristics of moving objects such as motion, position and shapes.

Formally, the preferred EM consists of a set of interdependent objects $O_i(t)$. This set in turn is comprised of a set of dynamic objects $D_{...}(t)$ and a set of static objects $S_{...}$. For instance, vehicles moving in a courtyard are dynamic objects; pillars standing in the courtyard are static objects. The time variance of the set $O_i(t)$ is a result of the time variation of the dynamic objects.

As befit their name, static objects do not vary with time. The set of values of these objects at any instant comprises the state of the system S(t). The preferred EM uses a layered model to represent objects at different levels of abstraction, such that there is a strong correlation between objects at different abstractions. Figure 4 shows some of the possible layers of the environment model, and how each layers communicates independently with other modules. Reference A. Katkere and R. Jain, A framework for information assimilation, in *Exploratory Vision* edited by M. Landy, et al., 1995.

To ensure consistency, any changes that occur in one level should be propagated to other levels (higher and lower), or at least tagged as an apparent inconsistency for future updating.

In general, propagation from higher to lower levels of abstractions is easier than vice versa. Accordingly, changes are attempted to be assimilated at as high level of abstraction as possible. Each dynamic object at the lowest level has a spatial extent of exactly one grid. Objects with higher extent are composed of these grid objects, and hence belong to higher levels.

SUBSTITUTE SHEET (RULE 26)

Direct information acquisition at higher levels must be followed by conversion of that information to the information at the densest level, so that information at all levels are consistent. It is important to come up with efficient access (and update) strategies at this level since this could potentially be the bottleneck of the entire representation and assimilation module.

Each dynamic object has several attributes, most basic being the confidence that it exists. Each of the above factors may contribute to either an increase or decrease in this confidence. These factors also affect the values of other object attributes. The value of an object $O_i(t)$, and hence, the state $S(t)$, may change due to the following factors: 1) New input information, i.e., new data regarding object position from the video data; 2) change in related model information; 3) advice from higher processes; and (4) decay (due to aging).

The preferred MPI-Video system provides facilities for managing dynamic and static objects, as is discussed further below in this section.

The EM, informed by the two-dimensional video data, provides a wealth of information not available from a single camera view. For instance, objects occluded in one camera view may be visible in another. In this case, comparison of objects in $D_{v,t}(t)$ at a particular time instant t with objects in $S_{v,t}$ can help anticipate and resolve such occlusions. The model, which takes inputs from both views, can continue to update the status of an object regardless of the occlusion in a particular camera plane. To maintain and manipulate information about position of static and dynamic objects in the environment, a representation must be chosen which facilitates maintenance of object positional information as well as supporting more sophisticated questions about object behavior. The preferred dynamic model relies on the following two components.

The first component is voxels. In this representation, the environment is divided up into a set of cubic volume elements, or voxels. Each voxel contains information such as which objects currently occupy this voxel, information about the history of objects in this voxel, an indication of which cameras can "see" this voxel. In this representation, objects can be described by the voxels they occupy. The voxel representation is discussed in greater detail in section 4.

The second component is (x,y,z) world coordinates. In this case, the environment and objects in the environment are represented using (x,y,z) world coordinates. Here objects can be

described by a centroid in (x,y,z), by bounding boxes, etc.

Each of these representations provides different support for modeling and data manipulation activities. The preferred MPI-Video system utilizes both representations.

13.2 Video Data Analysis and Information Assimilation

The Video Data Analyzer uses image and visual processing techniques to perform object detection, recognition and tracking in each of the camera planes corresponding to the different perspectives. The currently employed technique is based on differences in spatial position to determine object motion in each of the camera views. The technique is as follows.

First, each input frame is smoothed to remove some noise.

Second, the difference image $d_{...}$ is computed as follows. Only pixels that are in the focus of attention windows and that are not masked are considered. (Here F(t) refers to the pixels in the focus of attention, i.e., a region of interest in the frame t.)

$$d_{...} = \text{Threshold}\,(\text{Abs}(F_{...} - F_{.}), \text{threshold\_value}) \qquad (1)$$

To remove motion shadows, the following operation is done:

$$d_{.}^{*} = d_{...} \,\&\, d_{...} \qquad (2)$$

Third, components on the thresholded binary difference image are computed based on a 4-neighborhood criterion. Components that are too small or too big are thrown away as they usually constitute noise. Also frames that contain a large number of components are discarded. Both centroid (from first moments), and orientation and elongation (from the second moments) are extracted for each component.

Fourth, any of several optional filters can be applied to the components obtained from the previous step. These filters include, merging of overlapping bounding boxes, hard limits of orientation and elongation, distance from expected features etc.

The list of components associated with each camera is sent from the Video Analysis unit to the Assimilator module which integrates data derived from the multiple streams into a comprehensive representation of the environment.

The Assimilator module maintains a record of all objects in the environment. When new data arrives from the Video Data Analysis module the Assimilator determines if the new data corresponds to an object whose identity it currently maintains. If

so, it uses the new data to update the object information. Otherwise, it instantiates a new object with the received information. The following steps are employed to update objects.

First, the list of 2D object bounding boxes is further filtered based on global knowledge.

Second, the footprint of each bounding box is projected to the primary surface of motion by intersecting a ray drawn from the optic center of that particular camera through the foot of the bounding box with the ground surface.

Third, each valid footprint is tested for membership with existing objects and the observation is added as support to the closest object, if any. If no object is close enough, a new object hypothesis is created.

Fourth, all supporting observations are used (with appropriate weighting based on distance from the camera, direction of motion, etc.) to update the position of each object.

Fifth, the object positions are projected into the next frame based on a domain dependent tracker.

More sophisticated tracking mechanisms are easily integrated into the preferred system. A current area of our research seeks to employ additional methods to determine and maintain object identity. For instance, active contour models can be employed in each of the cameras to track object movements. See A. M. Baumberg and D. C. Hogg, An efficient method for contour tracking using active shape models, *Technical Report 94.11*, School of Computer Studies, University of Leeds, April, 1994. See also M. Kass, A. Witkin, and D. Terzopolous, Snakes: Active contour models, *International Journal of Computer Vision*, pages 321-331, 1988. See also F. Leymarie and M. D. Levine, Tracking deformable objects in the plane using an active contour model, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):617-634, June 1993. Such methods provide a more refined representation of object shape and dynamics.

One important assumption that is made is that the "static" world is known a priori and the only elements of interest in the video frames are the objects that undergo some type of change, e.g., a player running on a field. In addition, we introduce additional constraints by requiring cameras to be stationary and make the following realistic assumptions about objects of interest:

First, these objects are in motion most of the time.

Second, these objects move on known planar surfaces.

Third, these objects are visible from at least two viewpoints.

This knowledge of the "static" world is captured through the

camera calibration process which maps related locations in the two-dimensional video data to a fully three-dimensional representation of the world recorded by the cameras. If an event is seen in one camera, e.g., a wide receiver making a catch, or a dancer executing a jump, the system, using this mapping, can determine other cameras that are also recording the event, and where in the various video frames the event is occurring. Then a viewer, or the system, can choose between these different views of the action, subject to some preference. For example, the frames which provide a frontal view of the wide receiver or the dancer. This "best view" selection is described further below and in section 14.

When their positions and orientations are fixed, cameras can be calibrated before processing the video data using methods such as those described by Tsai and Lenz. See R. Y. Tsai and R. K. Lenz, A new technique for fully autonomous and efficient 3D robotics hand/eye calibration, *IEEE Transactions on Robotics and Automation*, 5(3):345-58, June 1989.

Calibration of moving cameras is a more difficult task and is currently an area of active research, e.g., ego motion. See E. S. Dickmanns and V. Graefe, Dynamic monocular machine vision, *Machine Vision and Applications*, 1:223-240, 1988.

The preferred MPI-Video system of the present invention has the capability to integrate these techniques into our analysis and assimilation modules when they become available. To date, evaluation of the preferred MPI-Video system has been done only by use of fixed cameras. The Assimilator maintains the Environment Model discussed above.

13.2.1 Camera Handoff

A key element in the maintenance of multiple camera views is the notion of a Camera Hand-off, here understood to be the event in which a dynamic object passes from one camera coverage zone to another. The Assimilator module also manages this processing task, maintaining a consistent representation of an object's identity and behavior during camera hand-off. This requires information about the object's position, its motion, etc.

Using the voxel information, noted above, we can determine which cameras can "see" (or partially "see") an object. Namely, a camera completely "sees" an object if all voxels occupied by the object are also seen by the camera. Let c(v) be the camera list, or set, associated with a particular voxel v, and V be the set of all voxels in which an object resides. Then, C_V is the complete

coverage, i.e. that set of cameras which can see all voxels in which an object resides and P_i is the partial coverage set, i.e. those cameras which can see some part of the object. These are defined as:

$$C_i = \bigcap_{v \in V} c(v) \quad (3)$$

$$P_i = \bigcup_{v \in V} c(v) \quad (4)$$

Thus, we can determine which cameras "see" a particular object by considering the intersection and/or union of the camera lists associated with the voxels in which the object resides. When an object moves between different zones of coverage, camera handoff is essentially automatic as a result of the a priori information regarding camera location and environment configuration. This is significant as it alleviates the necessity of reclassifying objects when they appear in a different camera view. That is, an object may enter a camera view and appear quite different then it did before, e.g., in this new perspective it may appear quite large.

However, reclassification is not necessary as the system, using its three dimensional model of the world, can determine which object this new camera measurement belongs to and can update the appropriate object accordingly. This capability is important for maintaining a temporally consistent representation of the objects in the environment. Such a temporal representation is necessary if the system is to keep track of object behavior and events unfolding in the environment over time.

13.3 Best View Selection

The View Selector can use a variety of criteria and metrics to determine a "best" view. Here, "best" is understood to be relative to a metric either specified by the user or employed by the system in one of its processing modules.

The best view concept can be illustrated by considering a case where there are N cameras monitoring the environment. Cameras will be denoted by C_i where the index $i \in \{1, \ldots, N\}$ varies over all cameras. At every time step, t, each camera produces a video frame, $F_{i,t}$. The term i_{sv}, will be used to indicate the best view index. That is, i_{sv} is the index of the camera which produces the best view at time t. Then, the best view is found by selecting the

frame from camera C_{iBV} at time t, i.e., the best view is $F_{iBV,t}$.

Some possible best view criteria include the least occluded view, the distance of the object to the camera, and object orientation.

In the case of a least occluded view criteria, the system chooses, at time t, that frame from the camera in which an object of interest is least occluded. Here, the best view camera index is defined according to the following criteria,

$$i_{BV} = \arg_i (\max S_i) \quad (5)$$

The object size metric S_i is given by:

$$S_i = \sum_{(x,y)} p(x, y) / S_{ei} \quad (6)$$

where $p(x, y) = 1$ if pixel $(x,y) \in R_i$ and 0 otherwise. R_i being the region of frame $F_{i,t}$ that contains the object of interest. We normalize the total size by the expected size, S_{ei} of the object, i.e., the number of pixels we expect the object to occupy in the camera view if no occlusion occurs. Finally, $\arg_i$ returns the index which optimizes this criteria.

In the case of an object distance of camera criteria, the best view is the frame in which an object of interest is closest to the corresponding camera.

$$i_{BV} = \arg_i (\min(D_i(t)) \quad (7)$$

where, $D_i(t)$ is the Euclidean distance between the (x,y,z) location of camera C_i and the world coordinates of the object of interest. The world coordinate representation, mentioned above, is most appropriate for this metric. Note also, that this criteria does not require any computation involving the data in the frames. However, it does depend on three-dimensional data available from the environment model.

For an orientation criteria a variety of possibilities exist. For instance, direction of object motion, that view in which a face is most evident, or the view in which the object is located closest to the center of the frame. This last metric is described by,

$$i_{BV} = \arg_i (\min (D_i(t)) \quad (8)$$

Here, $CD_i(t)$ is given by

$$CD_i(t) = \sqrt{(x(t) - (Xsize/2))^2 + (y,t) - (Ysize/2))^2} \quad (9)$$

The values Xsize and Ysize give the extent of the screen and (x(t), y(t)) are the screen coordinates of the object's two-dimensional centroid in frame $F_{i,t}$.

Combinations of metrics can also be employed. We can formulate a general representation of best view as follows:

$$i_{BV} = \arg_i (G(g_{i,t}(m_j(C_i)$$

$$| \; j \in \{1, \ldots, M\}, \; i \in \{1, \ldots, N\}; \; t \in \{1, \ldots, T\})))) \quad (10)$$

In this equation, each m_j is a metric, e.g., size as defined above, and we have M such metrics each of which is applied to the data from each camera, hence, the C_i terms in equation (10). Furthermore, each $g_{i,t}$ combines these metrics for C_i, e.g. as a weighted linear sum. The use of the time t in this equation supports a best view optimization which uses a temporal selection criteria involving many frames over time, as well as spatial metrics computed on each frame. This is addressed in the following paragraph. Finally, the criteria G chooses between all such combinations and $\arg_i$ selects the appropriate index. For instance, G might specify the minimum value.

For example, if we have three cameras (N = 3), two metrics (M = 2) and g specifying a linear weighted sum (using weights ω_1 and ω_2), G would pick the optimum of

$$g_{1,t} = \omega_1 m_1 (C_1) + \omega m_2 (C_1)$$

$$g_{2,t} = \omega_1 m_1 (C_2) + \omega m_2 (C_2)$$

$$g_{3,t} = \omega_1 m_1 (C_3) + \omega m_2 (C_3)$$

$$i_{BV} = \arg_i G(g_{1,t}, g_{2,t}, g_{3,t})$$

Again, G is a criteria which chooses the optimum from the set of gi;t's. Note that time does not appear explicitly in the right hand side of this equation, indicating that the same best view evaluation is applied at each time step t. Note, in this case, the same g (here, a weighted linear sum) is applied to all cameras, although, this need not be the case.

Two further generalizations are possible. Both are research issues we are currently addressing. Firstly, an optimization which

accounts for temporal conditions is possible. The best view is a frame from a particular camera. However, smoothness over time may also be important to the viewer or a system processing module. Thus, while a spatial metric such as object size or distance from a camera is important, a smooth sequence of frames with some minimum number of cuts (i.e. camera changes) may also be desired. Hence, best view selection can be a result of optimizing some spatial criteria such that a temporal criteria is also optimum.

A second generalization results if we consider the fact that the C_i's do not have to correspond to actual cameras views. That is, the preferred MPI-Video system has the capability of locating a camera anywhere in the environment. Thus, best view selection can be a function of data from actual cameras as well as from "virtual" cameras. In this case, equation 10 becomes a continuous function in the camera "index" variable. That is, we no longer have to restrict ourselves to the case of a finite number of cameras from which to chose the best view. Letting $\bar{x} = (x,y,z,\alpha,\beta,f)$ where (x,y,z) is the world coordinate position, or index, of the camera, α is a pan angle, β is camera tilt angle and f is a camera parameter which determines zoom in/out. The set of all such vectors $\bar{x}$ forms a 6-dimensional space, Ω. In Ω, (x,y,z) varies continuously over all points in R^3, $-\pi \leq \alpha, \beta \geq \pi$, and $f > 0$.

To determine the best view in the environment subject to some criteria, we search over all points in this space minimizing the optimization function. In this case, the best view is that camera positioned at location "x_{best}", where this value of the vector optimizes the constraint G given by:

$$G(g_{\bar{x},t}(m_j(\bar{x}) \mid j \in \{1,\ldots,M\}, t \in \{1,\ldots,T\}, \bar{x} \in \Omega)) \qquad (11)$$

The camera index, $\bar{x}$ can vary over all points in the environment and the system must determine, subject to a mathematical formulation of viewing specification, where to position the camera to satisfy a best view criteria. Views meeting this criteria can then be constructed using the techniques outlined in section 14.

For instance, using the same parameters as above, i.e. two metrics mi, the weighted linear summing function g and the criteria function G, we have the

$$g_{t\bar{x}} = \omega_1 m_1(\bar{x}) + \omega_2 m_2(\bar{x}) \qquad (12)$$

Then to determine the best view we find the value of $\bar{x}$ for

which

$$G(g_{cx}),\ \bar{x} \in \Omega \qquad (13)$$

is optimal.

Note that, assuming the computational power is available, the best view computation in equations (5), (7) and (8) can all be computed on the fly as video data comes into the system. More complex best view calculations, including those that optimize a temporal measure, may require buffered or stored data to perform best view selection.

Figure 23 shows how a selected image sequence is derived from four cameras and the determined "best" view. In this example, the "best" view is based upon two criteria, largest size and central location within the image where size takes precedence over location. Here, the function gi;tis just a simple weighted sum, as above, of the size and location metrics. The outlined frames represent chosen images which accommodate the selection criteria. Moreover, the oval tracings are superimposed onto the images to assist the viewer in tracking the desired object. The last row presents the preferred "best" view according to the desired criteria. In order to clarify the object's location, a digital zoom mechanism has been employed to the original image. In images T0 and T1, only from the view of camera 3 is the desired object visible. Although all camera views detect the object in image T2 and T3, the criteria selects the image with the greatest size. Once again in image T4, the object is only visible in camera 4.

13.4 Visualizer and Virtual View Builder

The visualizer and virtual view builder provides processing to create virtual camera views. These are views which are not produced by physical cameras. Rather they are realistic renditions composed from the available camera views and appear as if actually recorded. Such views are essential for immersive applications and are addressed in section 4 below.

13.5 Model and Analysis Interface

Figures 27, 28 and 29 show the current Motif-based preferred MPI-Video interface. This interface provides basic visualization of the model, the raw camera streams and the results of video data analysis applied to these streams. In addition, its menus provide control over the data flow as well as some other options. We are currently developing a hyper-media interface, in conjunction with

the development of a database system, which will extend the range of control and interaction a user has with the data input to and generated by our MPI-Video system. In the context of virtual scene creation such augmentations may include user selection of viewing position and manipulation (e.g. placement) of virtual model information into the environment.

The model shown in figures 27, 28 and 29 employs an (x, y, z) world coordinate, bounding box object representation. That is, the system tracks object centroid and uses a bounding box to indicate presence of an object at a particular location. A voxel-based representation supports finer resolution of object shape and location. Such a formulation is discussed in the next section 14.

14. Operation of Immersive Video (ImmV), or Interactive Telepresence, or Visual Reality (VisR)

Immersive and interactive telepresence is an idea that has captured the imagination of science fiction writers for a long time. Although not feasible in its entirety, it is conjectured that limited telepresence will play a major role in visual communication media in the foreseeable future. See, for example, N. Negroponte, *Being digital*, Knopf, New York, 1995.

In this section we describe Immersive Video (ImmV), a spatially-temporally realistic 3D rendition of real-world events. See the inventors' own papers: S. Moezzi, A. Katkere, S. Chatterjee, and R. Jain, Immersive Video, *Technical Report VCL-95-104*, Visual Computing Laboratory, University of California, San Diego, Mar. 1995; and S. Moezzi, A. Katkere, S. Chatterjee, and R. Jain, Visual Reality: Rendition of Live Events from Multi-Perspective Videos, *Technical Report VCL-95-102*, Visual Computing Laboratory, University of California, San Diego, Mar. 1995.

These events are simultaneously captured by video cameras placed at different locations in the environment. ImmV allows an interactive viewer, for example, to watch a broadcast of a football or soccer game from anywhere in the field, even from the position of the quarterback or "walk" through a live session of the US Congress.

Immersive Video involves manipulating, processing and compositing of video data, a research area that has received increasing attention. For example, there is a growing interest in generating a mosaic from a video sequence. See M. Hansen, P. Anandan, K. Dana, G. van der Wal, and P. Burt, Real-time Scene Stabilization and Mosaic Construction, in *ARPA Image Understanding*

Workshop, Monterey, CA, Nov. 13-16 1994. See also H. Sawhney, Motion Video Annotation and Analysis: An Overview, *Proc. 27th Asilomar Conference on Signals, Systems and Computers*, pages 85-89. IEEE, Nov. 1993.

The underlying task is to create larger images from frames obtained from a single-camera (panning) video stream. Video mosaicing has numerous applications including data compression. Another application is video enhancement. See M. Irani and S. Peleg, Motion analysis for image enhancement: resolution, occlusion, and transparency, *J. of Visual Communication and Image Representation*, 4(4):324-35, Dec. 1993. Yet another application is the generation of panoramic views. See R. Szeliski, Image mosaicing for tele-reality applications, *Proc. of Workshop on Applications of Computer Vision*, pages 44-53, Sarasota, FL, Dec. 1994. IEEE, IEEE Computer Society Press. See also L. McMillan. Acquiring Immersive Virtual Environments with an Uncalibrated Camera, *Technical Report TR95-006*, Computer Science Department, University of North Carolina, Chapel Hill, Apr. 1995. See also S. Mann and R. W. Picard. Virtual Bellows: Constructing High Quality Stills from Video. *Technical Report TR#259*, Media Lab, MIT, Cambridge, Mass., Nov. 1994. Still further applications included high-definition television, digital libraries etc.

To generate seamless video mosaics, registration and alignment of the frames from a sequence are critical issues. Simple, yet robust techniques have been suggested to alleviate this problem using multi-resolution area-based schemes. See M. Hansen, P. Anandan, K. Dana, and G. van der Wal et al., Real-time scene stabilization and mosaic construction, in *Proc. of Workshop on Applications of Computer Vision*, pages 54-62, Sarasota, FL, Dec. 1994. IEEE, IEEE Computer Society Press. For scenes containing dynamic objects, parallax has been used to extract dominant 2D and 3D motions which were then used in registration of the frames and generation of the mosaic. See H. Sawhney, S. Ayer, and M. Gorkani, Model-based 2D and 3D Dominant Motion Estimation for Mosaicing and Video Representation, *Technical report, IBM Almaden Res. Ctr.*, 1994.

For multiple moving objects in a scene, motion layers have been introduced where each dynamic object is assumed to move in a plane parallel to the camera. See J. Wang and E. Adelson, Representing moving images with layers, *IEEE Transactions on Image Processing*, 3(4):625-38, Sept. 1994. This permits segmentation of the video into different components each containing a dynamic object, which can then be interpreted and/or re-synthesized as a

video stream.

However, for immersive telepresence there is a need to generate 3D mosaics -- a "hyperMosaic" -- that can also handle multiple dynamic objects. Maintaining spatial-temporal coherence and consistency is integral to generation of such a HyperMosaic. In order to obtain 3D description, multiple perspectives that provide simultaneous coverage must therefore be used and their associated visual information integrated. Another necessary feature would be to provide a viewpoint that may be selected. The immersive video system and method of the present invention caters to these needs.

Immersive video requires sophisticated vision processing and modeling next described in Section 14.1. While Virtual Reality systems use graphical models and texture mapping to create realistic replicas of both static and dynamic components, in immersive video, distinctively, the data used is from actual video streams. This also aids in the rendition of exact ambiance, i.e. purely two dimensional image changes are also captured. For example, in ImmV, a viewer is able to move around a football stadium and watch the spectators from anywhere in the field and see them waving, moving, etc., in live video. For faithful reconstruction of realism, ImmV requires addressing issues such as synchronization between cameras, maintenance of consistency in both spatial and temporal signals, distributed processing and efficient data structures.

14.1 HyperMosaicing: Creating "Visual Realism"

Given the comprehensive model of the environment and accurate external and internal camera calibration information, compositing new vistas is accomplished by mosaicing pixels from the appropriate video streams. Algorithm 1 shown in Figure 31 outlines the steps involved. Algorithm 1 is the vista compositing algorithm. At each time instant, multiple vistas are computed using the current dynamic model and video streams from multiple perspective. For stereo, vistas are created from left and right cameras.

A basic element of this algorithmic process is a set of transformations between the model (or world) coordinate system W : $\{(x_w, y_w, z_w)\}$, the coordinate system of the cameras C : $\{(x_c, y_c, z_c)\}$ and the vista coordinate system V : $\{(x_v, y_v, z_v)\}$. For each pixel, $(x_v, y_v, d_v(x_v, y_v))$, on the vista the corresponding point, (x_w, y_w, z_w), is found in the world coordinate system using its depth value.

$$[x_w \; y_w \; z_w \; 1]^T = M_v \cdot [x_v \; y_v \; z_v \; 1]^T \qquad (14)$$

where M_v is the 4 x 4 homogeneous transformation matrix representing transformation between V and the world W [6].

This point is then projected onto each of the camera image planes c.

$$[x_c\ y_c\ z_c\ 1]^T = M_c^{-1} \cdot [x_w\ y_w\ z_w\ 1]^T \quad (15)$$

where M_c is the 4 x 4 homogeneous transformation matrix representing transformation between c and the world.

These points $(x_c, y_c, z_c)\ \forall$ c, are tested for occlusion from that view by comparing z_c with the depth value of the corresponding pixel. At this point, we have several candidates that could be used for the pixel (x_v, y_v) for the vista. Each candidate view cv is evaluated using the following two criteria:

First, the angle A subtended by line a of Figure 31 with the object point (x_w, y_w, z_w), computed using the cosine formula:

$$A = \arccos\left((b^2 + c^2 - a^2)\right) / (2bc) \quad (16)$$

See, for example, R. Courant and D. Hilbert, *Methods of Mathematical Physics*, volume 1. New York: Interscience Publishers, first english edition, 1953.

Second, the distance of the object point (x_w, y_w, z_w) from camera window coordinate (x_c, y_c), which is the depth value $d_c(x_c, y_c)$.

The evaluation criterion ecv for each candidate view cv is:

$$e_{cv} = f\ (A,\ B * d_c(x_c, y_c))\ ,\ \text{where B is a small number} \quad (17)$$

14.2 Immersive Video Prototype and Results

Our Immersive Video prototype is built on top of our MPI-Video system. See S. Chatterjee, R. Jain, A. Katkere, P. Kelly, D. Kuramura, and S. Moezzi, *Modeling and Interactivity in MPI-Video*, Technical Report VCL-94-104, Visual Computing Laboratory, UCSD, Dec. 1994; and A. Katkere, S. Moezzi, and R. Jain, Global Multi-Perspective Perception for Autonomous Mobile Robots, *Technical Report VCL-95-101*, Visual Computing Laboratory, UCSD, 1995.

People in the scene are detected and modeled as cylinders in our current implementation. For our experiments, a one minute long scene was digitized, at 6 frames/sec, from a half hour recording of four video cameras overlooking a typical campus scene. The digitized scene covers three pedestrians, a vehicle, and two bicyclists moving between coverage zones. Figure 22 shows the

relative placements of all four cameras. Frames from four cameras (for the same arbitrary time instant, 00:21:08:02) are shown in Figure 27. The scene contains three walkers. Note that though the zones of coverage have significant overlaps, they are not identical, thus, effectively increasing the overall zone being covered.

Some of the vistas generated by the prototype immersive video system of the present invention are shown in Figures 29a and 29b. White portions represent areas not covered by any camera. Note how each of the perspectives shown is completely different from any of the four original camera views.

Figure 29b illustrates how photo-realistic video images are generated by the system for a given viewpoint, in this case a ground level view overlooking the scene entrance. This view was generated by the prototype immersive video system using the comprehensive 3D model built by the MPI-Video modeling system and employing Algorithm 1 for the corresponding video frames shown in Figure 28. Note that this perspective is entirely different from the original views. A panoramic view of the same scene may also be produced. A bird's eye view of the walkway for the same time instant is shown in Figure 29a. Again, white portions represent areas not covered by any camera. Note the alignment of the circular arc. Images from all four cameras contributed towards the construction of the views.

Figure 28 also illustrates immersive abilities of the immersive video technology of the present invention by presenting selected frames from a 116-frame sequence generated for a walk through the entire courtyard. The walk through sequence illustrates how an event can be viewed from any perspective, while taking into account true object bearings and occlusions.

14.3 Discussion on the Representations

In this section 14, the concept for Immersive Video for rendition of live video from multiple perspectives has been described, and key aspects of the prototype system are described and shown. Although the system is at an early stage, it has been illustrated that immersive video can be achieved using today's technology and that photo-realistic video from arbitrary perspectives can be generated given appropriate camera coverage.

One of the limitations of the immersive video system, highlighted in closeups of people is simplistic modeling of dynamic objects (as bounding cylinders). While this simplification permitted development of a complete and fairly functional

prototype, such quirks should be, can be, and will be removed to achieve a greater degree of immersion. Towards this end, objects should be modeled more accurately. Two ways of achieving this are contemplated: detecting objects using predicted contours (Kalman snakes) and integrating these contours across perspectives, and using voxel-based integration. See D. Terzopolous and R. Szeliski, Tracking with Kalman snakes, in A. Blake and A. Yuille, editors, *Active Vision*, pages 3-20, MIT Press, Cambridge, Mass., 1992. See also D. Koller, J. Weber, and J. Malik, Robust Multiple Car Tracking with Occlusion Reasoning, *Proc. 3rd European Conference on Computer Vision*, pages 189-96, Stockholm, Sweden, May 1994. Springer-Verlag.

In the next section 15, how better object models can be built using voxels, and how this will improve the building of virtual vistas, is briefly described.

14.4 Voxel-Based Object Models

Voxels (or Spatial Occupancy Enumeration Cells) -- which are cells on a three-dimensional grid representing spatial occupancy -- provide one way of building accurate and tight object models. See J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practices*, Addison-Wesley Publishing Company, Inc., second edition, 1990.

Using techniques to determine occupancy of the voxels, the immersive video system of the present invention builds an accurate three dimensional model of the environment. An a priori static model (which occupies majority of filled space) is used to determine default occupancy of the voxels. To build the dynamic model, the occupancy of only those voxels whose state could have changed from the previous time instant is continuously determined. Using higher level knowledge, and information from prior processing, this computation may be, and preferably is, restricted to expected locations of dynamic objects.

The set of points that denote motion in an image can be computed using Algorithm 2 shown in Figure 32. Algorithm 2 is the voxel-construction-and-visualization-for-moving-objects algorithm.

This set subtends a portion of three dimensional space where motion might have occurred. Figure 23 and the diagrammatic portion of Figure 31 illustrate the viewing frustrums that define this space. Treating voxels as a accumulative array to hold positive and negative evidence of occupancy,the positive evidence of occupancy for this subtended space can be increased. Similarly, the space not subtended by motion points contribute to the negative

evidence. Assuming synchronized video streams, this information is accumulated over multiple perspectives (as shown in Figure 23 and the diagrammatic portion of Figure 31). A suitably selected threshold will separate voxels that receive positive support from multiple perspectives. Such voxels, with a high probability, represent dynamic objects. Algorithm 2 of Figure 32 shows the exact steps involved in this process.

Voxels are generated by integrating motion information across the four frames of Figure 27. The physical dimension of each voxel is 8 dm^3 or 2x2x2 dm^3. Comparing this with the cylindrical approximations of the MPI-Video modeling system, it is evident that more realistic virtual vistas can be created with voxels. Close contour approximations like Kalman snakes can also be used to achieve similar improvements.

14.4.1 Discussion on Computational and Storage Efficiency of Voxels

Voxels have been traditionally vilified for their extreme computing and storage requirements. To even completely fill a relatively small area like the courtyard used in the prototype system, some 14.4 million 1 dm^3 voxels are needed. With the recent and ongoing advances in storage and computing, this discussion may be moot. High speed, multi-processor desk-top machines with enormous amounts of RAM and secondary storage have arrived (e.g., high-end desk top computers from SGI). However for efficiency considerations and elegance, it is herein discussed how storage and computing requirements can greatly be reduced using certain assumptions and optimization.

One basic assumption is that motion is restricted to a small subset of the total three dimensional space and the static portion of the world is known a priori. Hence a combination of efficient geometry-based representation, like the Inventor format, can be used. See J. Wernecke, *The Inventor Mentor: Programming Object-Oriented 3D Graphics with Open T M Inventor*; Release 2. Addison-Wesley Publishing Company, 1994. Given that a three dimensional structure can be derived out of such a format, it is then necessary just model the dynamic portions using voxels.

Next, several assumptions are made about the dynamic objects:

First, the dynamic objects are assumed to be limited in their vertical extent. E.g., in the prototype immersive video system, all dynamic objects are in the range of 10-20 dm in height.

Second, bounds are put on where the objects may be at the current time instant based on prior state, tracking information,

assumptions about surfaces of motion etc.

The former assumption reduces the number of voxels by limiting the vertmcal dimension. Using the latter assumption, voxels are dynamically allocated to certain limited regions in the environment, and it is assumed that the remaining space retains the characteristics of the a priori static model. With this assumption, the number of voxels become a function of the number of expected dynamic objects instead of being a function of the total modeled space. While making these assumptions, and using two representations, slightly complicates spatial reasoning, the complexity in terms of storage and computation is greatly reduced.

In addition, to reduce the computational complexity of Algorithm 2, it is preferred to build look-up tables a priori to store the projection of each voxel on each camera. Since the relationship between each camera and the world is accurately known, this is a valid optimization.

15. Immersive Video / MPI-Video Prototype Implementation

This section provides some details on our MPI-Video prototype system used in the creation of the "virtual views" discussed in section 14.

Figure 15 shows the hardware configuration of the prototype immersive video system incorporating MPI video. The preferred setup consists of several independent heterogeneous computers. Ideally, one work station is used to process data from a single camera, preferably a Model 10 or 20 work station available from Sun. However, using a socket-based protocol multiple video processing modules can run on a reduced number of work stations (down to a single work station). In addition, a central (master) graphics work station (a SGI Indigo^2, Indy or Challenge) controls these video processing work stations (slaves) and maintains the Environment Model (and associated temporal database). The central master and the remote slaves communicate at a high symbolic level and minimal image information is exchanged. For instance, as will be discussed further below, object bounding box information is sent from the slaves to the master. Thus, actual image data need not be ex-changed, resulting in a very low required network bandwidth for master-slave communication. The work stations in the prototype system are connected on a 120 Mbps Ethernet switch which guarantees full-speed point-to-point connection.

The master-slave information exchange protocol is as follows:

First, the master initializes graphics, the database and the Environment Model (EM), and waits on a pre-specified port.

Second, based on its knowledge of the network, machine throughput etc. a separate process starts the slave processes on selected remote machines.

Third, each slave contacts the master (using pre-specified machine-port combination) and a initialization hand-shaking protocol ensues.

Fourth, the master acknowledges each slave and sends it initialization information, e.g., where the images are actually stored (for the laboratory case), the starting frame and frame interval, camera-specific image-processing information like thresholds, masks etc.

Fifth, each slave initializes itself based on the information sent by the master.

Sixth, once the initialization is completed, the master processes individual cameras as described in the next steps.

Seventh, whenever a frame from a specific camera needs to be processed the master sends a request to that particular slave with information about processing the frame viz. focus of attention windows frame specific thresholds and other parameters, current and expected locations and identifications of moving objects etc. and continues its processing (modeling and user interaction). (The focus of attention is essentially a region of interest in the image specifying where the visual processing algorithms should concentrate their action.) In synchronous mode, requests to all slaves are sent simultaneously and the integration is done after all slaves have responded. In asynchronous mode, this will not necessarily go in unison.

Eighth, when a reply is received, the frame information is used to update the Environment Model (EM). The following subsections present more detail on the individual components of our MPI-Video architecture. Virtual view synthesis is discussed in greater detail below.

16. Conclusions

Immersive Video so far presented has used multi-perspective video and a priori maps to construct three-dimensional models that can be used in interaction and immersion for diverse virtual world applications. One of these application is real-time virtual video, or virtual television, or telepresence -- next discussed in the following section 6. Various ways of presenting virtual video information have been discussed. Selection of the best view, creation of visually realistic virtual views, and interactive querying of the model have also been discussed. The actual

implementation of an immersive video system presented show that construction of video-based immersive environments is feasible and viable. The goal of the initial prototype immersive video system was not only to build a complete working system, but to also build a test-bed for the continuing development of more complicated and refined algorithms and techniques yet to be developed and tested. Towards this end, simple analysis and modeling techniques were used. Future work includes making these more sophisticated so that truly immersive environments can be constructed and used.

17. Immersive Telepresence

Immersive telepresence, or visual reality, is an immersive, interactive and realistic real-time rendition of real-world events captured by multiple video cameras placed at different locations in the environment. It is the real-time rendition of virtual video; "virtual television" instead of just "virtual video".

Unlike virtual reality, which is synthesized using graphical primitives, visual reality provides total immersion in live events. For example, a viewer can elect to watch a live broadcast of a football or soccer game from anywhere in the field. As with immersive video, immersive telepresence is based on and incorporates Multiple Perspective Interactive Video (MPI-Video) infrastructure for processing video data from multiple perspectives. In this section the particular adaptations of immersive video/MPI video for the implementation of immersive telepresence, or just plain "telepresence", are discussed. It is particularly shown and discussed as to how immersive telepresence may become an integral part of future television.

Alas, the drawings of this specification, being both (i) static, and (ii) two-dimensional, necessarily give only partial renditions of both (i) dynamic video and (ii) stereoscopy. Exemplary stereoscopic views produced by the immersive video system of the present invention respectively for the left and the right eyes are shown in Figures 14a, 14b and also 15a, 15b. In actual use both images are presented so as to be gated to an associated eye by such well-known virtual reality equipments as the "CrystalEyes" 3D Liquid Crystal Shutter (LCS) technology eyewear available from Stereographics Corporation.

It also impossible to convey in the drawings when something is happening in real time. In some cases the multiple video feeds from a scene that was processed in real time to present telepresence to a user/view were also recorded and were then later processed as immersive video. If the processing is the same then,

quite obviously, the presentations are also the same. Accordingly, some of the following discussion of exemplary results of immersive telepresence will refer to the same figures as did the discussion of immersive video!

The distinctions of note between immersive telepresence and immersive video are these. First, more computer processing time is clearly available in non-real time immersive video than in immersive telepresence. This may not be, however, of any great significance. More importantly, with immersive video the scene model may be revised, so as to improve the video renderings on an iterative basis and/or to account for scene occurrences that are unanticipated and not within the modeled space (e.g., the parachutist falling in elevation into the scene of a football game, which motion is totally unlike the anticipated motion of the football players and is not at or near ground level). The scene models used for immersive telepresence have been developed, and validated, for virtual video.

To be processed into immersive telepresence, it is not required that a scene should be "canned", or rote. It is, however, required that the structure of the scene (note that the scene has "structure", and is not a "windy jungle") should be, to a certain extent, pre-processed into a scene model. Therefore, not only does the scene model of a "football game" cover all football games, or the scene model of a "prizefight" cover all prizefights, but a scene model of a "news conference" may be pretty good at capturing the human actors therein, or a scene model of a "terrain scene including freeways from multiple helicopters" may be pretty good at capturing and displaying buildings, vehicles and pedestrians". The former two models are, of course, usable by sports broadcast organizations in the televising of scheduled events. However the last two models are usable by broadcast news organizations in the televising of events that may be unscheduled.

Competition by software developers in the development, and licensing, of scene models may arise. A television broadcaster able to obtain multiple television feeds would select and use the telepresence model giving best performance. Ultimately scene models will grow in sophistication, integration, and comprehensiveness, becoming able to do better in presentation, with fewer video feeds, faster.

17.1 The Use of Immersive Telepresence

It is conjectured that telepresence will play a major role in visual communication media. See N. Negroponte, *Being digital*.

Knopf, New York, 1995. Telepresence is generally understood in the context of Virtual reality (VR) with displays of real, remote scenes. This specification and this section instead describe immersive telepresence, being the real-time interactive and realistic rendition of real-world events, i.e., television where the viewer cannot control (does not interact with) what is happening in a real world scene, but can interact with how the scene is viewed.

Jaron Lanier defines Virtual Reality as an immersive, interactive simulation of realistic or imaginary environments. See J. Lanier. Virtual reality: the promise of the future. *Interactive Learning International*, 8(4):275-9, Oct.-Dec. 1992. The new concept called visual reality is an immersive, interactive and realistic rendition of real-world events simultaneously captured by video cameras placed at different locations in the environment. In contrast with virtual reality, or VR, where one can interact with and view a virtual world, visual reality, or VisR, permits a viewer/user one, for example, to watch a live broadcast of a football or soccer game from anywhere in the field, even from the position of the quarterback! Visual reality uses the Multiple Perspective Interactive Video (MPI-Video) infrastructure. See S. Chatterjee, R. Jain, A. Katkere, P. Kelly, D. Kuramura, and S. Moezzi, Modeling and interactivity in MPI-video, *Technical Report VCL-94-104, Visual Computing Lab, UCSD*, Dec. 1994.

MPI-Video is a move away from conventional video-based systems which permit users only a limited amount of control and insight into the data. Traditional systems provide a sparse set of actions such as fast-forward, rewind and play of stored information. No provision for automatic analysis and management of the raw video data is available.

Visual Reality involves manipulating, processing and compositing of video data, a research area that has received increasing attention. For example, there is a growing interest in generating a mosaic from a video sequence. See M. Hansen, P. Anandan, K. Dana, G. van der Wal, and P. Burt, Real-time scene stabilization and mosaic construction, in *ARPA Image Understanding Workshop*, Monterey, CA, Nov. 13-16 1994. See also H. Sawhney, Motion video annotation and analysis: An overview, in *Proc. 27th Asilomar Conference on Signals, Systems and Computers*, pages 85-89. IEEE, Nov. 1993.

The underlying task in video mosaicing is to create larger images from frames obtained as a video stream. Video mosaicing has numerous applications including data compression, video

enhancement. See M. Irani and S. Peleg, Motion analysis for image enhancement: resolution, occlusion, and transparency, in *J. of Visual Communication and Image Representation*, 4(4):324-35, Dec. 1993. See also R. Szeliski, Image mosaicing for tele-reality applications, in *Proc. of Workshop on Applications of Computer Vision*, pages 44-53, Sarasota, FL, Dec. 1994. See also the IEEE, IEEE Comput. Soc. Press. high-definition television, digital libraries etc.

To generate video mosaics, registration and alignment of the frames from a sequence are critical issues. Simple, yet robust techniques have been suggested to alleviate this problem using multi-resolution area-based schemes. See M. Hansen, P. Anandan, K. Dana, and G. van der Wal et al., Real-time scene stabilization and mosaic construction, in *Proc. of Workshop on Applications of Computer Vision*, pages 54-62, Sarasota, FL, Dec. 1994. IEEE, IEEE Comput. Soc. Press. For scenes containing dynamic objects, parallax has been used to extract dominant 2-D and 3-D motions, which were then used in registration of the frames and generation of the video mosaic. See H. Sawhney, S. Ayer, and M. Gorkani, Model-based 2D and 3D dominant motion estimation for mosaicing and video representation, Technical Report, *IBM Almaden Res. Ctr.*, 1994.

For multiple moving objects in a scene, motion layers have been introduced where each dynamic object is assumed to move in a plane parallel to the camera. See J. Wang and E. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(4):625-38, Sept. 1994. This permits segmentation of the video into different components each containing a dynamic object, which components can then be interpreted and/or re-synthesized as a video stream.

However, for immersive telepresence there is a need to generate a comprehensive 3-D mosaic that can handle multiple dynamic objects as well. The name affixed to this process is "hyper-mosaicing". In order to obtain a 3-D description, multiple perspectives that provide simultaneous coverage must be used, and their associated visual information must be integrated. Another necessary feature is provide a selected viewpoint. Visual reality satisfies all these requirements.

These issues, and a description of a prototype visual reality, are contained in the following sections. Section 6.2 recapitulates the concepts of MPI-Video as is especially applied to VisR. Section 6.3 provides implementation details and present results for the same campus walkway covered by multiple video cameras -- only

this time as television in real time as opposed to non-real-time video. Future directions for VisR are outlined in section 6.4.

17.2 Visual Reality using Multi-Perspective Videos

Visual Reality requires sophisticated vision processing, as well as modeling and visualization. Some of these are readily available under MPI-Video. See S. Chatterjee, R. Jain, A. Katkere, P. Kelly, D. Kuramura, and S. Moezzi. Modeling and interactivity in MPI-video. *Technical Report VCL-94-104, Visual Computing Lab, UCSD,* Dec. 1994. MPI-Video is a framework for management and interactive access to multiple streams of video data capturing different perspectives of related events. It involves automatic or semi-automatic extraction of content from the data streams, modeling of the scene observed by these video streams, management of raw, derived and associated data. These video data streams can reflect different views of events such as movements of people and vehicles. In addition, MPI-Video also facilitates access to raw and derived data through a sophisticated hypermedia and query interface. Thus a user, or an automated system, can query about objects and events in the scene, follow a specified object as it moves between zones of camera coverage and select from multiple views. A schematic showing multiple camera coverage typical in a MPI-Video analysis was shown in Figure 22a.

For a true immersive experience, a viewer should be able to view the events from anywhere. To achieve this, vistas composed from appropriate video streams must be made available. Generating these vistas requires a comprehensive three-dimensional model that represents events captured from these multiple perspective videos. Given multiple 'static' views, it is possible theoretically to extract this 3-D model using low-level vision algorithms e.g., shape from X methods. However, it is widely accepted that current methods make certain assumptions that cannot be met and that are, in general, non-robust. For environments that are mostly static, a priori information, e.g. a CSG/CAD model of the scene, can be used in conjunction with camera information to bypass the extraction of the static portions and to reduce the complexity of processing the dynamic portions. This is analogous to extracting the optical flow in only the portions of the scene where brightness changes are expected due to motion (flow discontinuities). This is exploited in the present implementation of Visual Reality (VisR) to create realistic models.

While in virtual reality (VR) texture mapping is used to create realistic replicas of both static and dynamic components, in

SUBSTITUTE SHEET (RULE 26)

visual reality (VisR), distinctively, actual video streams ar used. Ideally, exact ambiance will always be reflected in the rendition, i.e., purely two dimensional images changes are also captured. For example, in VisR a viewer is able to move around a football stadium and watch the spectators from anywhere in the field and see them waving, moving, etc.

17.3 Approach and Results

The current prototype immersive telepresence system is used in conjunction with multiple actual video feeds of a real-world scene to compose vistas of this scene. Experimental results obtained for a campus scene show how an interactive viewer can 'walk through' this dynamic, live environment in as it exists in real time (e.g., as seen through television).

17.3.1 Building a Comprehensive, Dynamic 3-D Model

Any comprehensive three-dimensional model consists of static and dynamic components. For the static model a priori information e.g., a CAD model, about the environment is used. The model views are then be registered with the cameras. Accurate camera calibration plays a significant role in this. For the dynamic model, it is necessary to (i) detect the objects in the images from different views, (ii) position them in 3-D using calibration information, (iii) associate them across multiple perspectives, and (iv) obtain their 3-D shape characteristics. These issues hereinafter next described are also accorded explanation in the technical report by S. Chatterjee, R. Jain, A. Katkere, P. Kelly, D. Kuramura, and S. Moezzi titled Modeling and interactivity in MPI-video, *Technical Report VCL-94-104, Visual Computing Lab, UCSD,* Dec. 1994. See also A. Katkere, S. Moezzi, and R. Jain, Global multi-perspective perception for autonomous mobile robots, Technical Report VCL-95-101, *Visual Computing Laboratory*, UCSD, 1995.

It is widely accepted that if a 3-D model of the scene is available, then many of the low-level processing tasks can be simplified. See Y. Roth and R. Jain, Simulation and expectation in sensor-based systems, *International Journal of Pattern Recognition and Artificial Intelligence*, 7(1):145-73, Feb. 1993.
For example, associating images taken at different times or from different views becomes easier if one has some knowledge about the 3-D scene points and the camera calibration parameters (both internal and external). In VisR this is exploited -- as it was in immersive video -- to simplify vision tasks, e.g., segmentation

etc. (model-based vision).

In the approach of the present invention cameras are assumed to be calibrated a priori. Using pre-computed camera coverage tables 2-D observations are mapped into 3-D model space and 3-D expectations into 2D image space. Note the bi-directional operation.

For the prototype VisR system, a complete, geometric 3-D model of a campus scene was built using architectural map data. At a basic level, the VisR system must and does extract information from all the video streams, reconciling extracted information with the 3-D model. As such, a scene representation was chosen which facilitates maintenance of object's location and shape information.

In the preferred VisR, or telepresence, system, object information is stored as a combination of voxel representation, grid-map representation and object-location representation. Note the somewhat lavish use of information. The systems of the present invention are generally compute limited, and are generally not limited in storage. Consider also that more and faster storage may be primarily a matter of expending more money, but there is a limit to how fast the computers can compute no matter how much money is expended. Accordingly, it is generally better to maintain an information-rich texture from which the computer(s) can quickly recognize and maintain scene objects than to use a more parsimonious data representation at the expense of greater computational requirements.

For each view, the prototype VisR, or telepresence, system is able to compute the 3-D position of each dynamic object detected by a motion segmentation module in real time. A priori information about the scene and camera calibration parameters, coupled with the assumption that all dynamic objects move on planar surfaces permits object detection and localization. Note the similarity in constraints to object motion(s), and the use of a priori information, to immersive video. Using projective geometry, necessary positional information is extracted from each view. The extracted information is then assimilated and stored in a 2D grid representing the viewing area.

17.3.1.1 Dynamic Objects

While more sophisticated detection, recognition and tracking algorithms are still susceptible of development and application, the initial prototype VisR, or immersive telepresence, system uses simple yet robust motion detection and tracking. Connected components labelling is used on the difference images to detect

moving objects. This also initializes/updates a tracker which exchanges information with a global tracker that maintains state information of all the moving objects.

Even though instantaneous 3-D shape information is not currently processed due to lack of computation power, it is an option under development. See A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models, in *Proc. Workshop on Motion of Non-rigid and Articulated Objects*, pages 194-9, Austin, TX, Nov. 1994, IEEE, Comput. Soc. Press. Video processing is simplified by "focus of attention rectangles" and pre-computed static mask images delineating portions of a camera view which cannot possibly have any interesting motion. The computation of the former is done using current locations of the object hypotheses in each view and projected locations in the next view. The latter is created by painting out areas of each view not on the planar surface (walls, for example).

17.2 Vista Compositing

Given the comprehensive model the environment and accurate external and internal camera calibration information, compositing new vistas at the view-port is simply a number of transformations between the model (or world) coordinate system (x_w, y_w, z_w), the coordinate system of the cameras (x_c, y_c, z_c) and the view-port coordinate system (x_v, y_v, z_v). Each pixel (on the composited display) is projected on the world coordinate system. The world point is then projected onto each of the camera image planes and tested for occlusion from that view. Given all such un-occluded points (and their intensity values), the following selection criteria is used. First, the pixel value for the point which subtends the smallest angle with respect to the vista and is closest to the viewing position is used in the rendition. This is then repeated for every time instant (or every frame) assuming stationary view-port. To generate a "fly-by" sequence this is repeated for every position of the view-port in the world coordinate. Note that this also makes the task of handling sudden zonal illumination changes ("spotlight effects") easier. Algorithm 1 shown in Figure 31 outlines the steps involved. Note that the generation of panoramic views from any view-port is a by product with a suitable selection of camera parameters (angle of view, depth of field etc.).

17.3 Visual Reality Prototype and Results

The prototype application of the immersive telepresence system

of the present invention involved the same campus scene (actually, a courtyard) as was used for the immersive video. The scene was covered by four cameras at different locations. Figure 22a shows the model schematic (of the environment) along with the camera positions. Note that though the zones of camera coverage have significant overlaps, they are not identical, thus, effectively increasing the overall zone being covered.

To illustrate the compositing effect, cameras with different physical characteristics were used. To study the dynamic objects, people were allowed to saunter through the scene. Although in the current version, no articulated motion analysis is incorporated, work is underway to integrate such and other higher-order behaviors. See S. Niyogi and E. Adelson. Analyzing gait with spatio-temporal surfaces, in *Proc. of Workshop on Motion of Non-Rigid and Articulated Objects*, pages 64-9, Austin, TX, Nov. 1994, IEEE, Comput. Soc. Press.

As previously discussed, Figure 27 shows corresponding frames from four views of the courtyard with three people walking. The model view of the scene is overlaid on each image. Figure 28 shows some "snapshots" from a 116-frame sequence generated for a "walk through" the entire courtyard. People in the scene are detected and modeled as cylinders in our current implementation as shown in Figure 29. The "walk" sequence illustrates how an event can be viewed from anywhere, while taking into account true object bearings and pertinent shadows. Also as previously discussed, Figure 29a shows a ground level view of the scene, and Figure 29b a bird's eye from the top of the scene. Each view is without correspondence to any view within any of the video streams.

17.4 Conclusions and Future Work

The prototype VisR system serves to render live video from multiple perspectives. This provides a true immersive telepresence with simple processing modules. The incorporation of more sophisticated vision modules, e.g., detecting objects using predicted contours (Kalman snakes), distributed processing of the video streams, etc., is expected in the future.

In the prototype system each of the cameras is assumed to be fixed with respect to the static environment. An incorporation of camera panning and zooming into the model is expected to be useful in representing sporting events. To date no problems with camera jitter, frame dropouts etc. have been encountered in the prototype system. However, if the frame digitalizations are synchronized, then any such occurrences as these can be handled quite

SUBSTITUTE SHEET (RULE 26)

efficiently.

Given the nature of events transpiring in the scene, and the simplified processing transpiring, digitalization in the prototype system was set at 6 frames/second. This can be easily made adaptive for each individual camera.

The next generation of television is anticipated to contain features of VisR, although a great deal of work remains in either reducing or meeting some of the stringent computational and memory demands. See N. Negroponte, *Being digital*, Knopf, New York, 1995.

18. Immersive Video/Television At the Present Time, or How to Use Five Hundred Television Channels Beneficially

The diverse sophisticated video presentations discussed in this specification are so discussed in the necessarily formative terminology of the present time, when not enough people have seen these effects of these video presentations so as to give them the popular names that they will, no doubt, ultimately assume. Moreover, the showing within this specification of examples of these video presentations is limited to drawings that are both (i) static and (ii) one dimensional (and, as will be explained, are of scenes intentionally rendered sufficiently crudely so that certain effects can be observed). According to the limits of description and of illustration, it is perhaps difficult for the reader to know what is reality and what is "hype", and what can be done right now (circa 1995) versus what is likely coming in the future world of video and television. The inventors endeavor to be candid, and blunt, while acknowledging that they cannot perfectly foresee the future.

Immersive video may be divided into real-time applications, i.e., immersive television, and all other, non-real-time, applications where there is, mercifully, more time to process video of a scene.

Both applications are presently at developed to a usable, and arguably a practically usable, state. Each application is, however, perceived to have a separate development and migration path, roughly similarly as video and television entertainments constitute a separate market from computer games and interactive computerized tutors at the present time.

18.1 Monitoring Live Events in Real Time or Near Real Time

With high speed video digitalizers, an immersive video system based on a single engineering work station class computer can, at the present time, process and monitor (being two separate things)

the video of live events in real time or near real time.

Such a system can, for example, perform the function of a "television sports director" -- at least so far as a "video sports director" focused on limited criteria -- reasonably well. The immersive video "sports director" would, for example, be an aid to the human sports director, who would control the live television primary feed of a televised sporting event such as a football game. The immersive video "sports director" might be tasked, for example, to "follow the football". This view could go out constantly upon a separate television channel.

Upon incipient use of an immersive video system so applied, however, the view would normally only be accessed upon selected occasions such as, for example, an "instant replay". The synthesized virtual view is immediately ready, without any such delay as normally presently occurs while humans figure out what camera or cameras really did show the best view(s) of a football play, upon the occasion of an instant replay. For example, the synthesized view generally presenting the "football" at center screen can be ordered. If a particular defensive back made a tackle, then his movements throughout the play may be of interest. In that case a sideline view, or helmet view, of this defensive back can be ordered.

With multiple computers, multiple video views can be simultaneously synthesized, each transmission upon a separate television channel. Certain channels would be devoted to views of certain players, etc.

As the performance of computer hardware and communication links increase, it may ultimately be possible to have television views on demand.

Another presently-realizable real-time application is security, as at, for example, airports. An immersive video system can be directed to synthesize and deliver up "heads-up facial view" images of people in a crowd, one after the next and continuously as and when camera(s) angle(s) permit the capture/synthesis of a quality image. Alternatively, the immersive system can image, re-image and synthetically image anything that its classification stage suspects to be a "firearm". Finally, just as the environment model of a football game expects the players to move but the field to remain fixed, the environment model of a secured area can expect the human actors therein to move but the moveable physical property (inventory) to remain fixed or relatively fixed, and not to merge inside the human images as might be the case if the property was being concealed for purposes of theft.

It will be understood that the essence of an immersive video system is image synthesis and presentation, and not image classification. However, by "forming up" images from desired optimal vantage points, and by operating under an environment model, the immersive video system has good ability (as it should, at its high cost) to permit existing computer image classification programs to successfully recognize deviations -- objects in the scene or events in the scene. Although human judgment as to what is being represented, and "seen", by the system may ultimately be required, the system, as a machine, is tireless and continuously regards the world that it views with an "attentiveness" not realizable by humans.

It should further be considered that the three-dimensional database, or world model, within an immersive video system can be the input to three-, as opposed to two-, dimensional classification programs. Human faces (heads) in particular might be matched against stored data representing existing, candidate, human heads in three dimensions. Even when humans regard "mug shots", they prefer both frontal and side views. Machine classification of human facial images, as just one example, is expected to be much improved if, instead of just one video view at an essentially random view angle, video of an entire observed head is available for comparison.

The ultimate use of real-time and near-real-time immersive video may in fact be in machine perception as opposed to human entertainment. The challenge of satisfying the military requirement of an autonomous vehicle that navigates in the environment, let alone the environment of a battlefield, is a very great one. The wondrous "visual world view" presented to our brains by our eyes is actually quite limited in acuity, sensitivity, spectral sensitivity, scale, detection of temporal phenomena, etc., etc. However, a human does a much better job of making sense of the environment than does a computer that may actually "see" better because the human's understanding, or "environmental model", of the real-world environment is much better than that of the computer. Command and control computers should perhaps compensate for the crudity of their environmental models by assimilating more video data inputs derived from more spatial sites. Interestingly, humans, as supported by present-day military computer systems, already recognize the great utility of sharing tactical information on a theater of warfare basis. In particular, the Naval Tactical Data System (NTDS) -- now almost forty years old -- permits sharing of the intelligence data developed from many

separate sensor platforms (ships, planes, submarines, etc.).

It may be essential that computers that operate autonomously or semi-autonomously during warfare should be allowed to likewise share and assimilate sensor information, particularly including video data, from multiple spatially separated platforms. In other words, although one robot tank seeing a battlefield from just one vantage point (even with binocular vision) may become totally lost, three or four such tanks together sharing information might be able to collectively "make sense" of what is going on. The immersive video system of the present invention is clearly involved with world-, or environment-, level integration of video information taken from spatially separated video sources (cameras), and it would be a mistake to think that the only function of an immersive video system is for the entertainment or education of humans.

An attached appendix contains the computer program source code for realizing immersive video in accordance with the present invention.

18.1 Processing of Video in Non-Real-Time

Meanwhile to developments in immersive television, the processing of video information -- which is not required to transpire in real time -- and the communication of video information -- which may be by disc or like transportable storage media instead of over land cable or radio frequency links -- may proceed in another direction. Anything event or scene that people wish to view with great exactitude, or to interact with realistically (which are not the same thing), can be very extensively "worked up" with considerable computer processing. A complete 3-D database of fine detail can be developed, over time and by computer processing, from historical multiple video feeds of anything from a football game to a stage play or, similarly to the more exotic scenes common in "surround vision" theaters, travel locales and action sequences. When recorded, a scene from the 3-D database can be "played back" at normal, real-time, speeds and in accordance with the particular desires of a particular end viewer/user by use of a computer, normally a personal computer, of much less power than the computer(s) that created the 3-D database. Every man or woman will thus be accorded an aid to his or her imagination, and can, as did the fictional Walter Mitty, enter into any scene and into any event.

For example, one immediate use of immersive video is in the analysis of athlete behaviors. An athlete, athlete in training, or aspiring athlete performs a sports motion such as, for example, a

golf swing that is videotaped from multiple, typically three, camera perspectives. A 3-D video model of the swing, which may only be a matter of ten or so seconds, is constructed at leisure, perhaps over some minutes in a personal computer. A student golfer and/or his/her instructor can subsequently play back the swing from any perspective that best suits observation of its salient characteristics, or those of its attributes that are undergoing corrective revision. If two such 3-D models of the same golfer are made, one can be compared against the other for deviations, which may possibly be presented as colored areas or the like on the video screen. If a model or an expert golfer, or a composite of expert golfers, is made, then the swing of the student golfer can be compared in three dimensions to the swing(s) of the expert golfer(s).

Another use of machine-aided comparison, and content-based retrieval, of video, or video-type, images is in medicine. New generations of Magnetic Resonance Imaging (MRI) sensors are already poised to deliver physiological information in stereoscopic representation, for example as a 3-D model of the patient's brain facilitating the planning of neurosurgery. However, immediate medical applications of immersive video in accordance with the present invention are much more mundane. A primary care physician might, instead of just recording patient height and weight and relying on his or her memory from one patient visit to the next, might simply videotape the standing patient's unclothed body from multiple perspectives at periodic intervals, an inexpensive procedure conducted in but a few seconds. Three-dimensional patient views constructed from each session could subsequently be compared to note changes in weight, general appearance, etc.

In the long term, the three-dimensional imaging of video information (which video information need not, however, have been derived from video cameras) as is performed by the immersive video system of the present invention will likely be useful for machine recognition of pathologies. For somewhat the same reasons that it is difficult for the computerized battlefield tank discussed above to find its way around on the battlefield from only a two-dimensional view thereof, a computer is inaccurate in interpreting, for example, x-ray mammograms, because it looks at only a two-dimensional image with deficient understanding of how the light and shadow depicted thereon translates to pathology of the breast. It is now so much that a tumor might be small, but that a small object shown at low contrast and high visual signal-to-noise is difficult to recognize in two dimensions. It is generally easier to

recognize, and to classify, a medical image in three dimensions because most of our bodies and their ailments -- excepting the skin and the retina -- are substantially three-dimensional.

Another use of the same 3-D human images realized with immersive video system of the present invention would be in video representations of the prospective results of reconstructive or cosmetic (plastic) surgery, or of exercise regimens. The surgeon or trainer would modify the body image, likely by manipulation of the 3-D image database as opposed to 2-D views thereof, much in the manner that any computerized video image is presently edited. The patient/client would be presented with the edited view(s) as being the possible or probable results of surgery, or of exercise.

In accordance with these and other possible variations and adaptations of the present invention, the scope of the invention should be determined in accordance with the following claims, only, and not solely in accordance with that embodiment within which the invention has been taught.

CLAIMS

What is claimed is:

1. A method of presenting a particular two-dimensional video image of a real-world three dimensional scene to a viewer comprising:

imaging in multiple video cameras each at a different spatial location multiple two-dimensional images of a real-world scene each at a different spatial perspective;

combining in a computer the multiple two-dimensional images of the scene into a three-dimensional model of the scene;

receiving in a the computer from a prospective viewer of the scene a viewer-specified criterion relative to which criterion the viewer wishes to view the scene;

producing in the computer from the three-dimensional model a particular two-dimensional image of the scene in accordance with the received viewer criterion; and

displaying in a video display the particular two-dimensional image of the real-world scene to the viewer.

2. The method according to claim 1 wherein the producing in the computer comprises:

selecting from the three-dimensional model a two-dimensional image corresponding to one of the images of the real-world scene that is imaged by one of the multiple video cameras.

3. The method according to claim 1 wherein the producing in the computer comprises:

synthesizing from the three-dimensional model a two-dimensional image that is without exact correspondence to any of the images of the real-world scene that are imaged by any of the multiple video cameras.

4. The method according to claim 1

wherein the receiving is of the viewer-specified criterion of a particular spatial perspective, relative to which particular spatial perspective the viewer wishes to view the scene; and

wherein the producing in the computer from the three-dimensional model is of a particular two-dimensional image of the scene in accordance with the particular spatial perspective criterion received from the viewer; and

wherein the displaying in the video display is of the

particular two-dimensional image of the scene that is in accordance with the particular spatial perspective received from the viewer.

5. The method according to claim 4 wherein the producing in the computer comprises:

selecting from the three-dimensional model an actual image of the scene as was imaged by a one of the multiple video cameras, this selected image being an actual image of the scene, out of all the actual images of the scene as were imaged by all the multiple video cameras, that is most closely in accordance with the particular spatial perspective criterion received from the viewer.

6. The method according to claim 5

wherein the selecting from the three-dimensional model is, over time, of plural actual images of the scene as are imaged, over time, by plural ones of the multiple video cameras;

wherein the computer does not invariably select from the three-dimensional model an image that arises from one only of the multiple video cameras, but instead selects plural images as arise over time from plural ones of the multiple video cameras.

7. The method according to claim 4 wherein the producing in the computer comprises:

synthesizing from the three-dimensional model a virtual image that is without correspondence to any of the images of the scene that are imaged by any of the multiple video cameras, this synthesized virtual image being in accordance with the particular spatial perspective criterion received from the viewer.

8. The method according to claim 1

wherein the combining is so as generate a three-dimensional model of the scene in which model objects in the scene are identified;

wherein the receiving is of the viewer-specified criterion of a selected object that the viewer wishes to particularly view within the scene; and

wherein the producing in the computer from the three-dimensional model is of a particular two-dimensional image of the selected object in the scene; and

wherein the displaying in the video display is of the particular two-dimensional image of the scene showing the viewer-selected object.

9. The method according to claim 8 wherein the viewer-selected object in the scene is static, and unmoving, in the scene.

10. The method according to claim 8 wherein the viewer-selected object in the scene is dynamic, and moving, in the scene.

11. The method according to claim 8 wherein the viewer selects the object that he or she wishes to particularly view in the scene by act of positioning a cursor on the video display, which cursor unambiguously specifies an object in the scene by an association between the object position and the cursor position in three dimensions and is thus a three-dimensional cursor.

12. The method according to claim 1

wherein the combining is so as generate a three-dimensional model of the scene in which model events in the scene are identified;

wherein the receiving is of the viewer-specified criterion of a selected event that the viewer wishes to particularly view within the scene; and

wherein the producing in the computer from the three-dimensional model is of a particular two-dimensional image of the selected event in the scene; and

wherein the displaying in the video display is of the particular two-dimensional image of the scene showing the viewer-selected event.

13. The method according to claim 12 wherein the viewer selects the event that he or she wishes to particularly view in the scene by act of positioning a cursor on the video display, which cursor unambiguously specifies an event in the scene by an association between the event position and the cursor position in three dimensions and is thus a three-dimensional cursor.

14. The method according to claim 1 performed in real time as television presented to a viewer interactively in accordance with the viewer-specified criterion.

15. A method of synthesizing a virtual video image from real video images obtained by a multiple real video cameras, the method comprising:

storing in a video image database the real two-dimensional video images of a scene from each of a multiplicity of real video

cameras;

creating in a computer from the multiplicity of stored two-dimensional video images a three-dimensional video database containing a three-dimensional video image of the scene; and

generating a two-dimensional virtual video image of the scene from the three-dimensional video database.

16. The method according to claim 15 wherein the generating comprises:

selecting from the three-dimensional video database a two-dimensional virtual video image of the scene that corresponds to a real two-dimensional video image of a scene.

17. The method according to claim 15 wherein the generating comprises:

synthesizing from the three-dimensional video database a two-dimensional virtual video image of the scene that is without correspondence to any real two-dimensional video image of a scene.

18. The method according to claim 15 that, between the creating and the generating, further comprises:

selecting a spatial perspective, which spatial perspective is not that of any of the multiplicity of real video cameras, on the scene as is imaged within the three-dimensional video database;

wherein the generating of the two-dimensional virtual video image is so as to show the scene from the selected spatial perspective.

19. The method according to claim 18 wherein the selected spatial perspective is static, and fixed, during the video of the scene.

20. The method according to claim 18 wherein the selected spatial perspective is dynamic, and variable, during the video of the scene.

21. The method according to claim 18 wherein the selected spatial perspective is so dynamic and variable dependent upon occurrences in the scene.

22. The method according to claim 15 that, between the creating and the generating, further comprises:

locating a selected object in the scene as is imaged within

the three-dimensional video database;

wherein the generating of the two-dimensional virtual video image is so as to best show the selected object.

23. The method according to claim 15 that, between the creating and the generating, further comprises:

dynamically tracking the scene as is imaged within the three-dimensional video database in order to recognize any occurrence of a predetermined event in the scene;

wherein the generating of the two-dimensional virtual video image is so as to best show the predetermined event.

24. The method according to claim 15 wherein the generating is of a selected two-dimensional virtual video image, on demand.

25. The method according to claim 15 wherein the generating of the selected two-dimensional virtual video image is in real time on demand, thus interactive virtual television.

26. A method of telepresence, being a video representation of being at real-world scene that is other than the instant scene of the viewer, the method comprising:

capturing video of a real-world scene from each of a multiplicity of different spatial perspectives on the scene;

creating from the captured video a full three-dimensional model of the scene;

producing from the three-dimensional model a video representation on the scene that is in accordance with the desired perspective on the scene of a viewer of the scene, thus immersive telepresence because the viewer can view the scene as if immersed therein, and as if present at the scene, all in accordance with his/her desires;

wherein the representation is called immersive telepresence because it appears to the viewer that, since the scene is presented as the viewer desires, the viewer is immersed in the scene;

wherein the viewer-desired perspective on the scene, and the video representation in accordance with this viewer-desired perspective, need not be in accordance with any of the captured video.

27. The method of immersive telepresence according to claim 26

wherein the video representation is in accordance with the

position and direction of the viewer's eyes and head, and exhibits motional parallax;

wherein motional parallax is, normally and conventionally, a three-dimensional effect where different views on the scene are produced as the viewer moves position even should the viewer have but one eye, making the viewer's brain to comprehend that the viewed scene is three-dimensional.

28. The method of immersive telepresence according to claim 26

wherein the video representation is stereoscopic;

wherein stereoscopy is, normally and conventionally, a three-dimensional effect where each of the viewer's two eyes sees a slightly different view on the scene, making the viewer's brain to comprehend that the viewed scene is three-dimensional even should the viewer not move his/her head or eyes in spatial position.

29. A method of telepresence, being a video representation of being at real-world scene that is other than the instant scene of the viewer, the method comprising:

capturing video of a real-world scene from a multiplicity of different spatial perspectives on the scene;

creating from the captured video a full three-dimensional model of the scene;

producing from the three-dimensional model a video representation on the scene that is in accordance with a predetermined criterion selected from among criteria including a perspective on the scene, an object in the scene and an event in the scene, thus interactive telepresence because the presentation to the viewer is interactive in accordance with the criterion;

wherein the video presentation of the scene in accordance with the criterion need not be in accordance with any of the captured video.

30. The method of viewer-interactive telepresence according to claim 29

wherein the video representation is in accordance with a criterion selected by the viewer, thus viewer-interactive telepresence.

31. The method of viewer-interactive telepresence according to claim 30 wherein the presentation is in accordance with the position and direction of the viewer's eyes and head, and exhibits

motional parallax.

32. The method of viewer-interactive telepresence according to claim 30 wherein the presentation exhibits stereoscopy.

33. An immersive video system for presenting video images of a real-world scene in accordance with a predetermined criterion, the system comprising:

a knowledge database containing information about the scene:

multiple video sources each at a different spatial location for producing multiple two-dimensional video images of a real-world scene each at a different spatial perspective;

a viewer interface at which a prospective viewer of the scene may specify a criterion relative to which criterion the viewer wishes to view the scene;

a computer, receiving the multiple two-dimensional video images of the scene from the multiple video cameras and the viewer-specified criterion from the viewer interface, the computer including

a video data analyzer for detecting and for tracking objects of potential interest and their locations in the scene,

an environmental model builder for combining multiple individual video images of the scene to build a three-dimensional dynamic model of the environment of the scene within which three-dimensional dynamic environmental model potential objects of interest in the scene are recorded along with their instant spatial locations, and

a viewer criterion interpreter for correlating the viewer-specified criterion with the objects of interest in the scene, and with the spatial locations of these objects, as recorded in the dynamic environmental model in order to produce parameters of perspective on the scene, and

a visualizer for generating, from the three-dimensional dynamic environmental model in accordance with the parameters of perspective, a particular two-dimensional video image of the scene; and

a video display, receiving the particular two-dimensional video image of the scene from the computer, for displaying this particular two-dimensional video image of the real-world scene to the viewer as that particular view of the scene which is in satisfaction of the viewer-specified criterion.

34. The immersive video system according to claim 33 wherein

the knowledge database contains data regarding at least two of
the geometry of the real-world scene,
potential shapes of objects in the real-world scene,
dynamic behaviors of objects in the real-world scene, and
a camera calibration model.

35. The immersive video system according to claim 33 wherein the knowledge database contains data regarding each of
the geometry of the real-world scene,
potential shapes of objects in the real-world scene,
dynamic behaviors of objects in the real-world scene, and
a camera calibration model.

36. The immersive video system according to claim 33 wherein the camera calibration model of the knowledge database includes at least one of
an internal camera calibration model, and
an external camera calibration model.

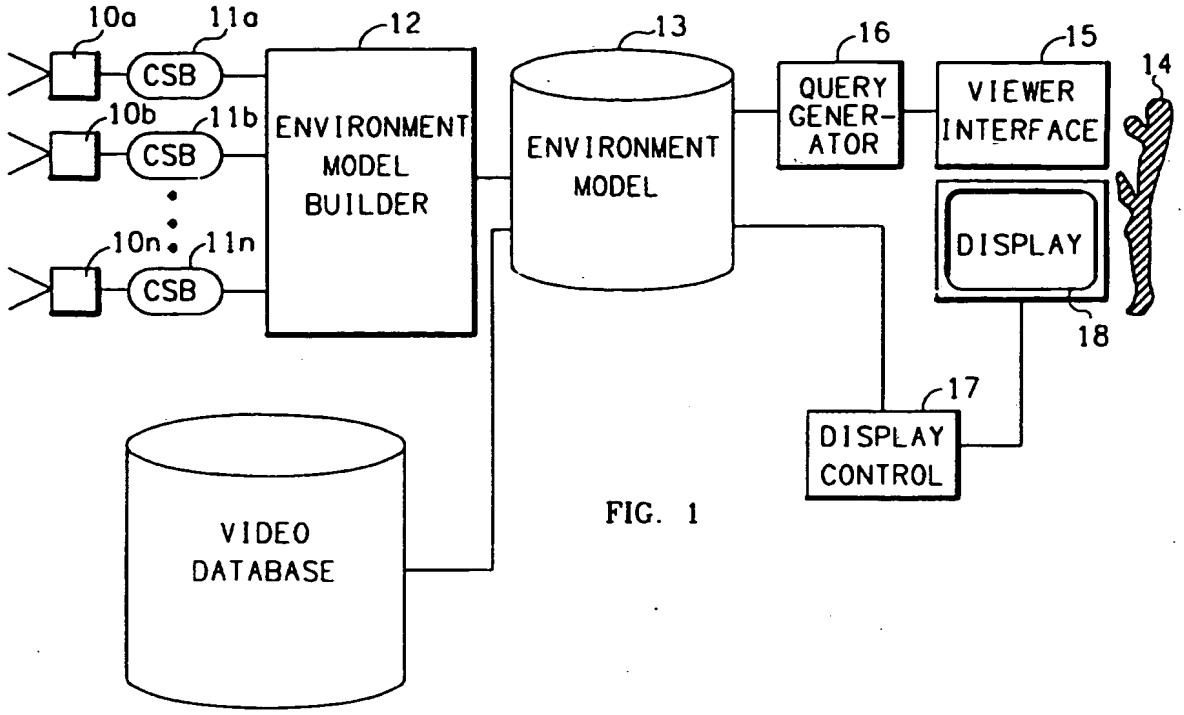

10a
11a
12
13
16
15
14
CSB
10b
11b
ENVIRONMENT MODEL BUILDER
ENVIRONMENT MODEL
QUERY GENER-ATOR
VIEWER INTERFACE
CSB
10n
11n
DISPLAY
18
17
CSB
DISPLAY CONTROL
VIDEO DATABASE

FIG. 1

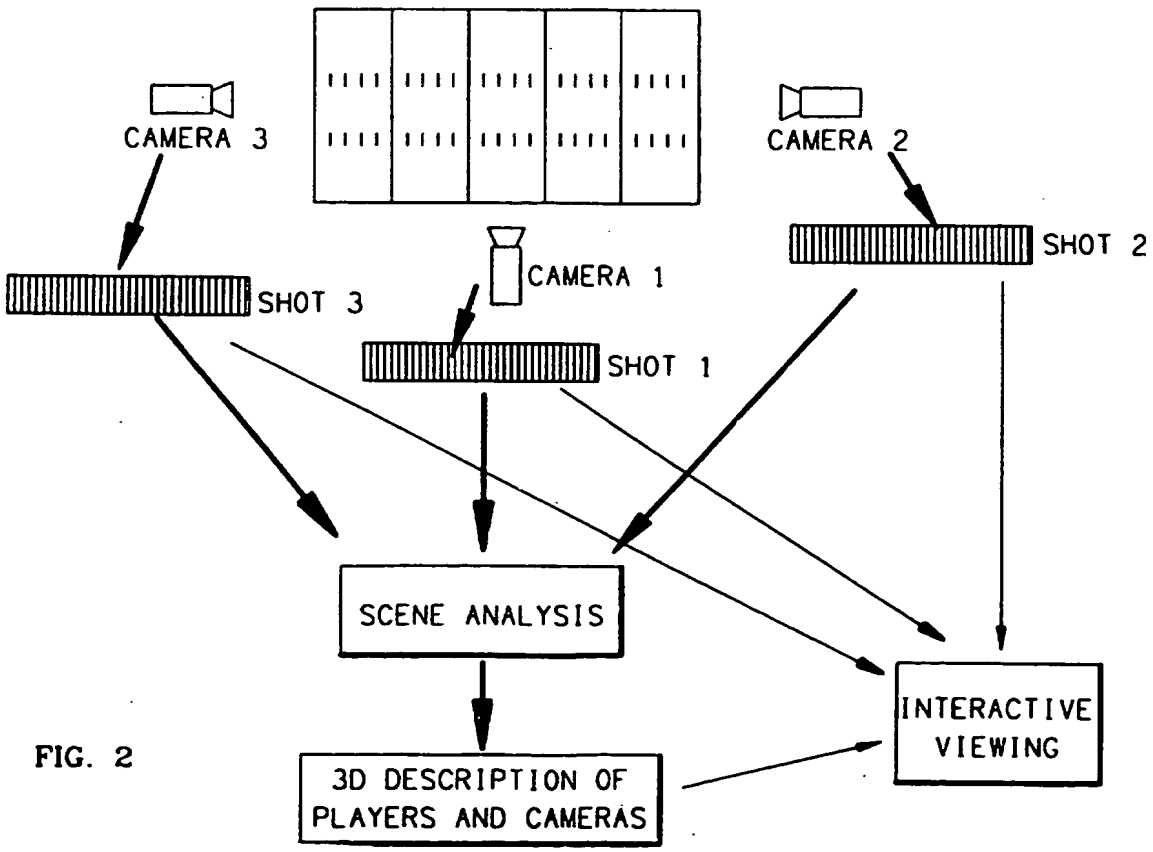

CAMERA 3
CAMERA 2
SHOT 2
CAMERA 1
SHOT 3
SHOT 1
SCENE ANALYSIS
INTERACTIVE VIEWING
3D DESCRIPTION OF PLAYERS AND CAMERAS

FIG. 2

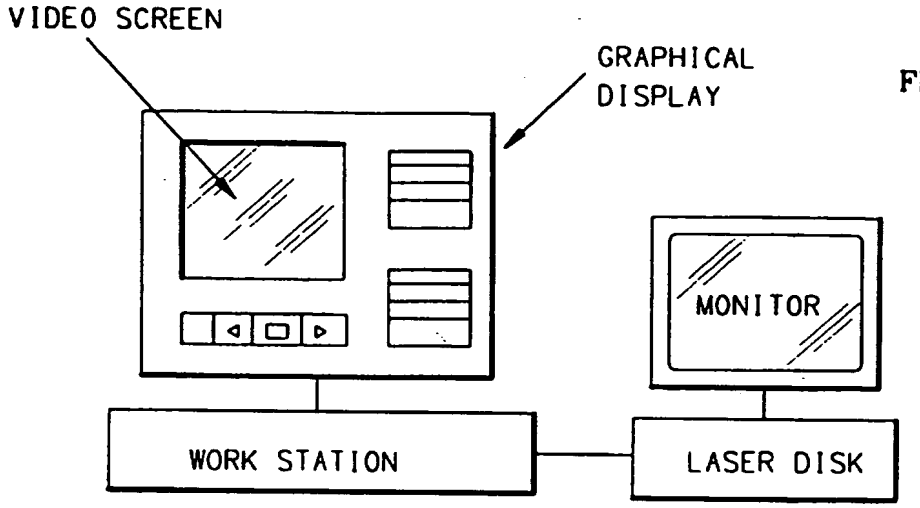

VIDEO SCREEN
GRAPHICAL DISPLAY
MONITOR
WORK STATION
LASER DISK

FIG. 3

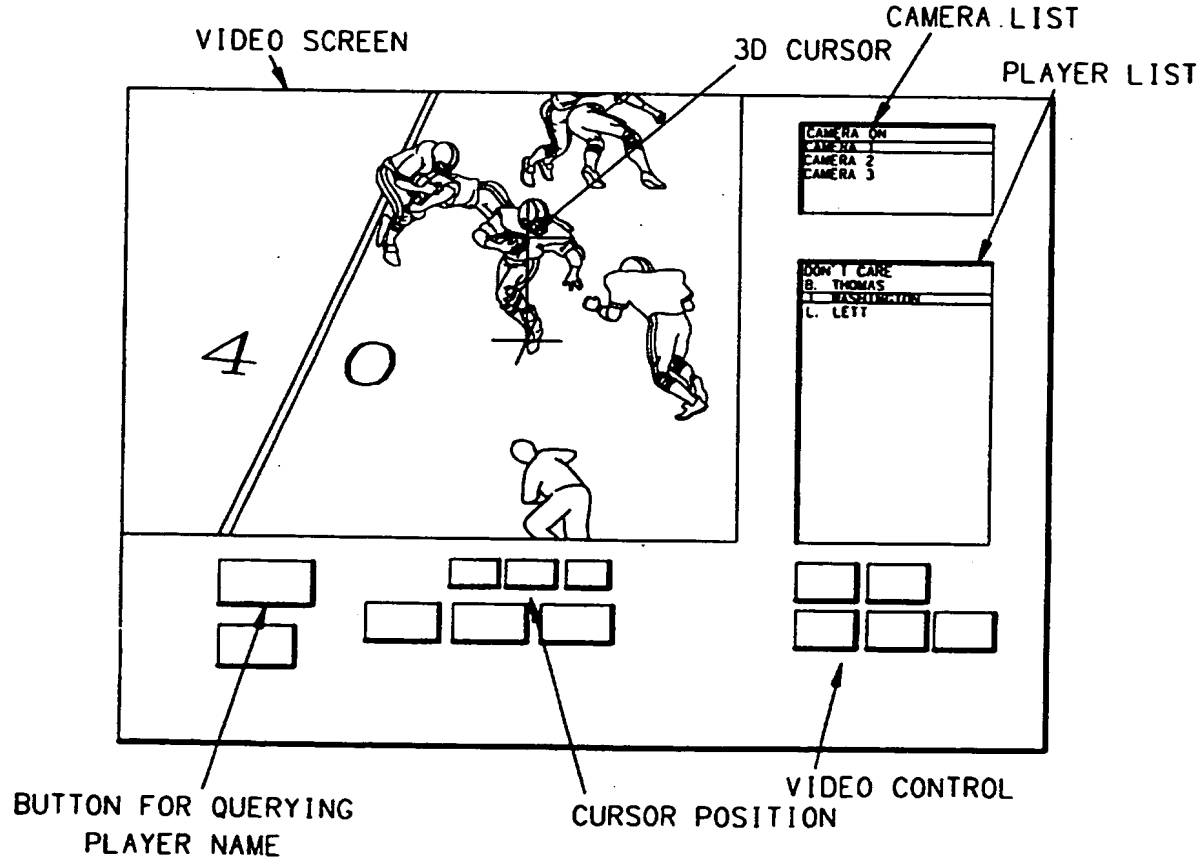
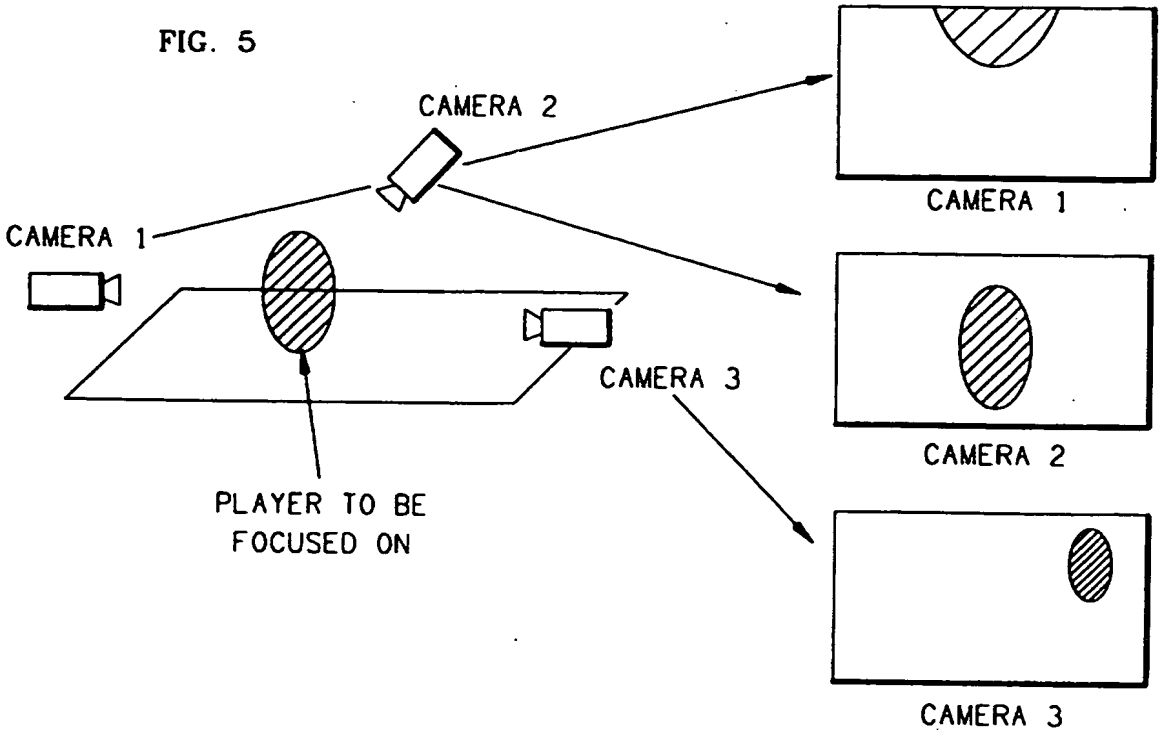

VIDEO SCREEN
3D CURSOR
CAMERA LIST
PLAYER LIST
CAMERA ON
CAMERA 2
CAMERA 3
DON'T CARE
B. THOMAS
L. LETT
BUTTON FOR QUERYING PLAYER NAME
CURSOR POSITION
VIDEO CONTROL

FIG. 4

FIG. 5


CAMERA 2
CAMERA 1
CAMERA 1
CAMERA 3
CAMERA 2
PLAYER TO BE FOCUSED ON
CAMERA 3

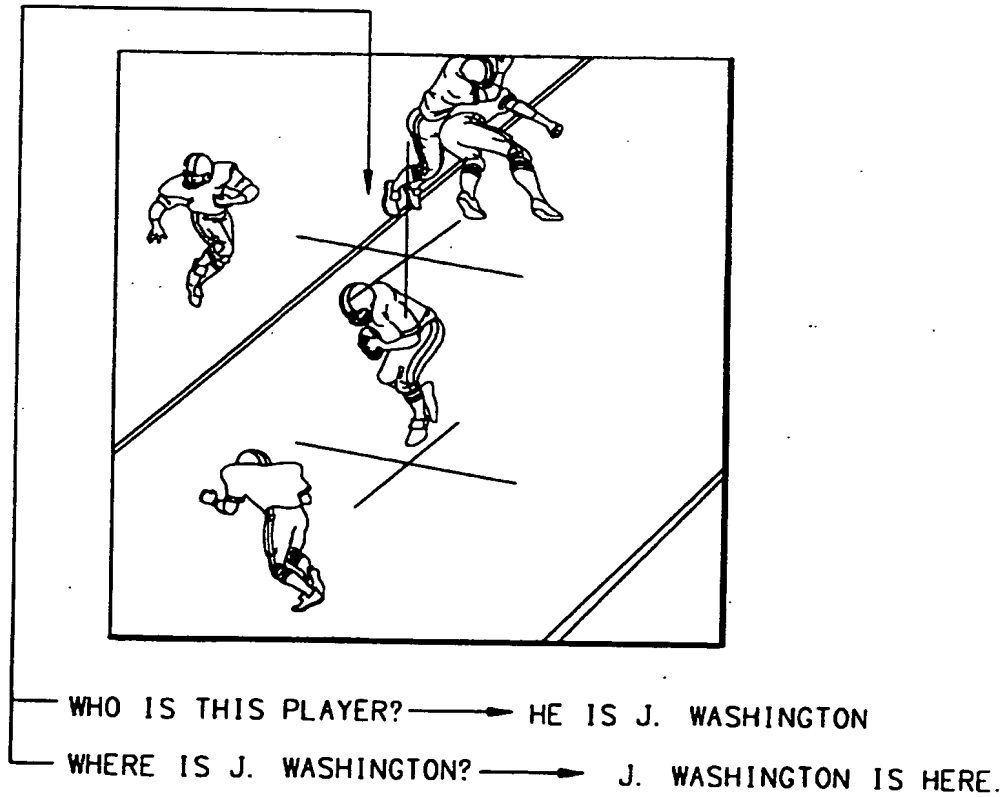
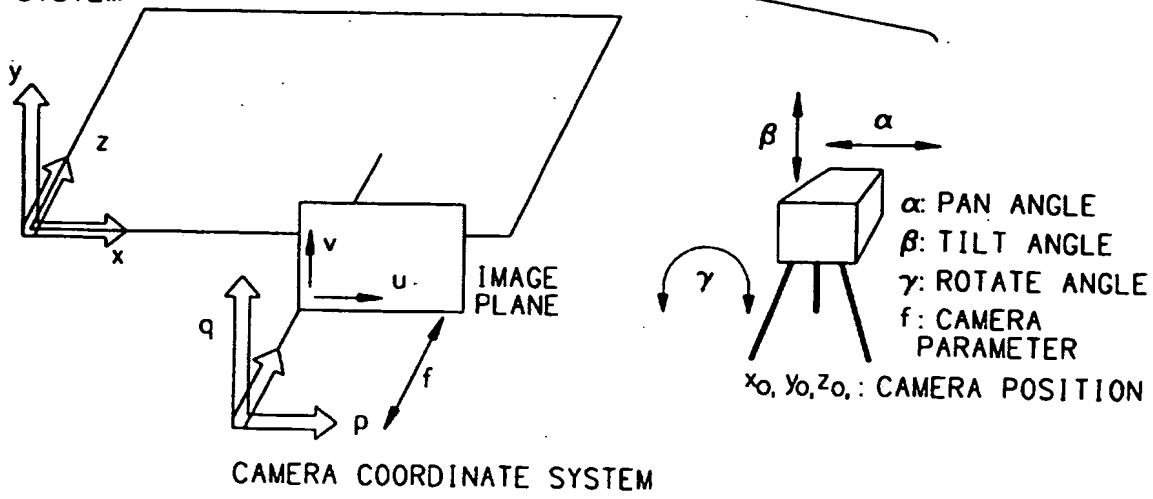

WHO IS THIS PLAYER? → HE IS J. WASHINGTON
WHERE IS J. WASHINGTON? → J. WASHINGTON IS HERE.

FIG. 6

FIG. 7


WORLD COORDINATE SYSTEM
y
z
x
v
u
IMAGE PLANE
q
p
f
CAMERA COORDINATE SYSTEM
β
α
γ
α: PAN ANGLE
β: TILT ANGLE
γ: ROTATE ANGLE
f: CAMERA PARAMETER
x_o, y_o, z_o: CAMERA POSITION

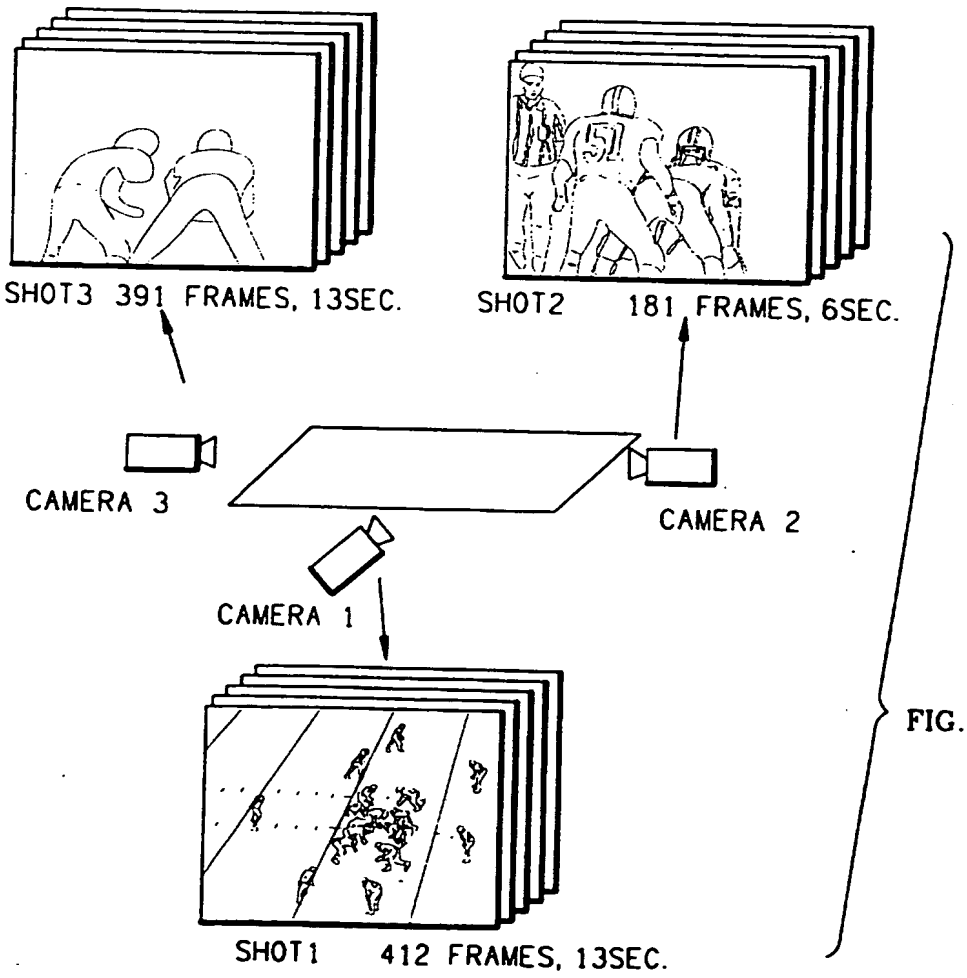

SHOT3 391 FRAMES, 13SEC.
SHOT2 181 FRAMES, 6SEC.
CAMERA 3
CAMERA 2
CAMERA 1
SHOT1 412 FRAMES, 13SEC.

FIG. 8

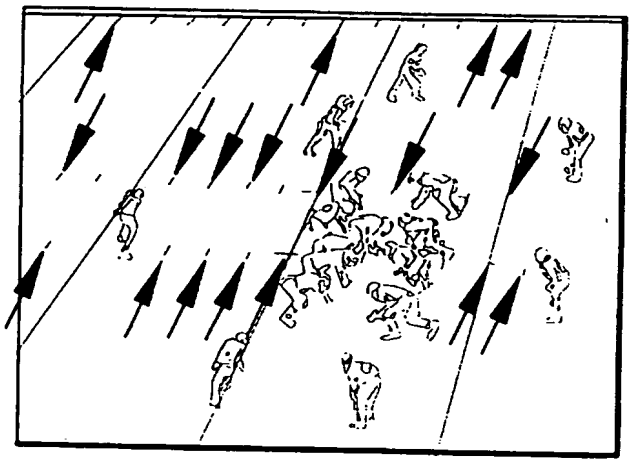


FIG. 9a

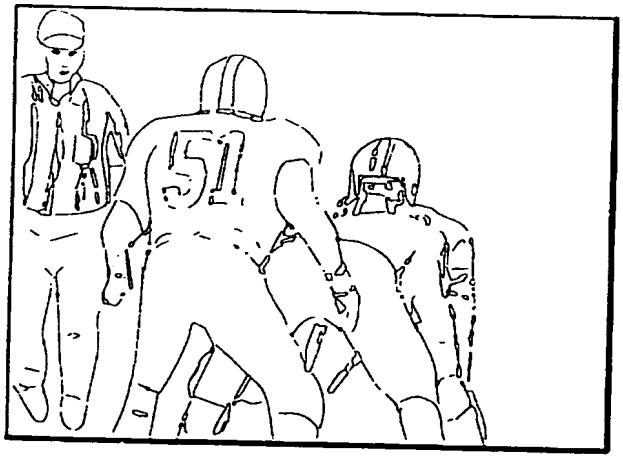
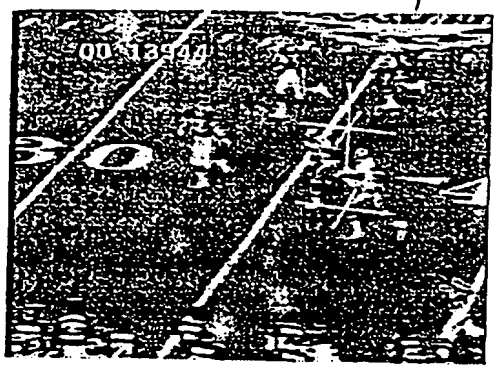


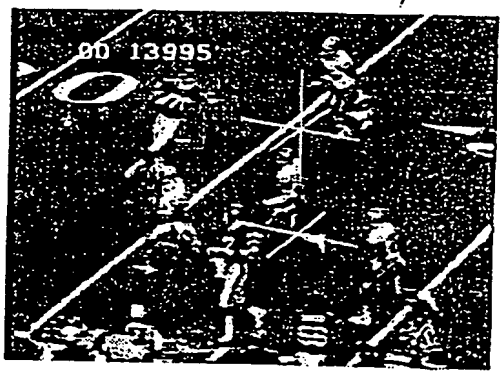
FIG. 9b

Fig. 16a

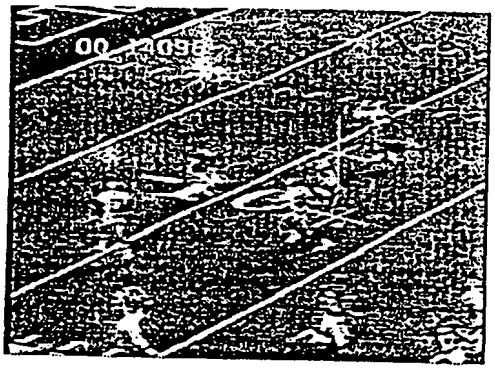


(36.9, 10.5)

Fig. 16b



(27.2, 6.1)



(13.1, 12.7)

Fig 16c

(x, z): cursor position
x: yardage from goal line
z: yardage from side line

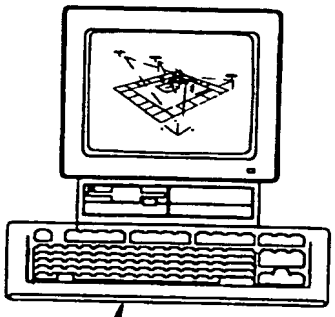


FIG. 11b

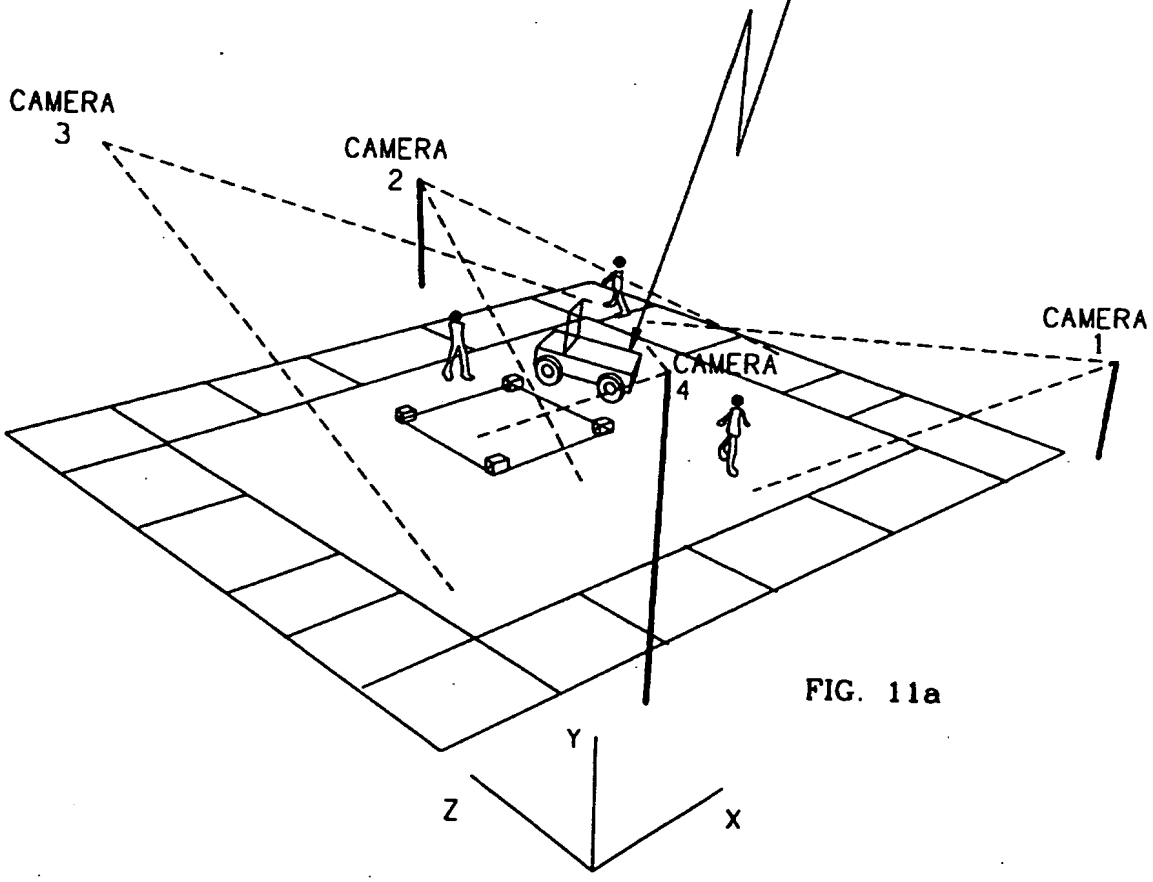


FIG. 11a

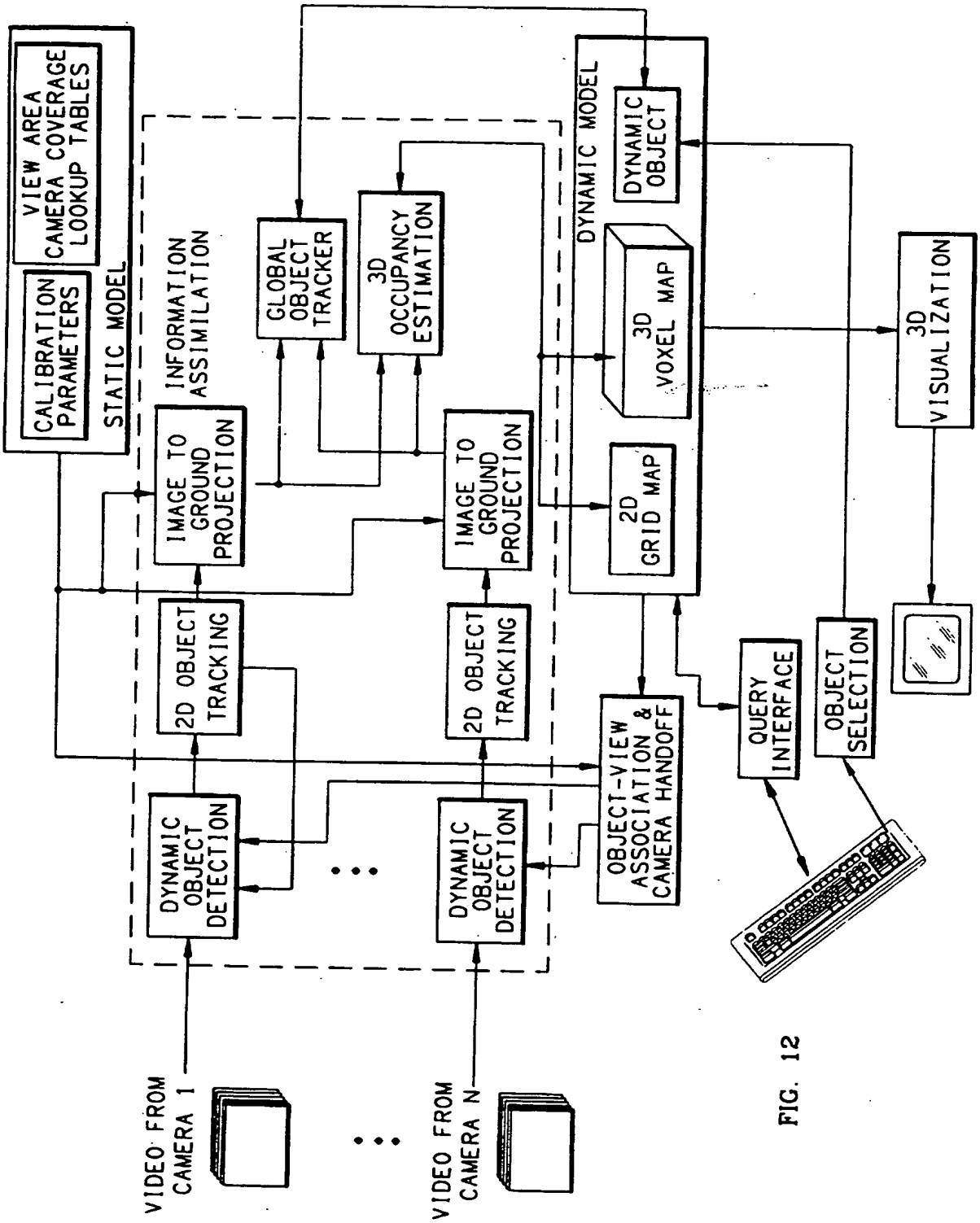

VIEW AREA CAMERA COVERAGE LOOKUP TABLES
CALIBRATION PARAMETERS
STATIC MODEL
INFORMATION ASSIMILATION
GLOBAL OBJECT TRACKER
3D OCCUPANCY ESTIMATION
DYNAMIC MODEL
DYNAMIC OBJECT
3D VOXEL MAP
3D VISUALIZATION
IMAGE TO GROUND PROJECTION
IMAGE TO GROUND PROJECTION
2D GRID MAP
2D OBJECT TRACKING
2D OBJECT TRACKING
QUERY INTERFACE
OBJECT SELECTION
DYNAMIC OBJECT DETECTION
DYNAMIC OBJECT DETECTION
OBJECT-VIEW ASSOCIATION & CAMERA HANDOFF
VIDEO FROM CAMERA 1
VIDEO FROM CAMERA N

FIG. 12

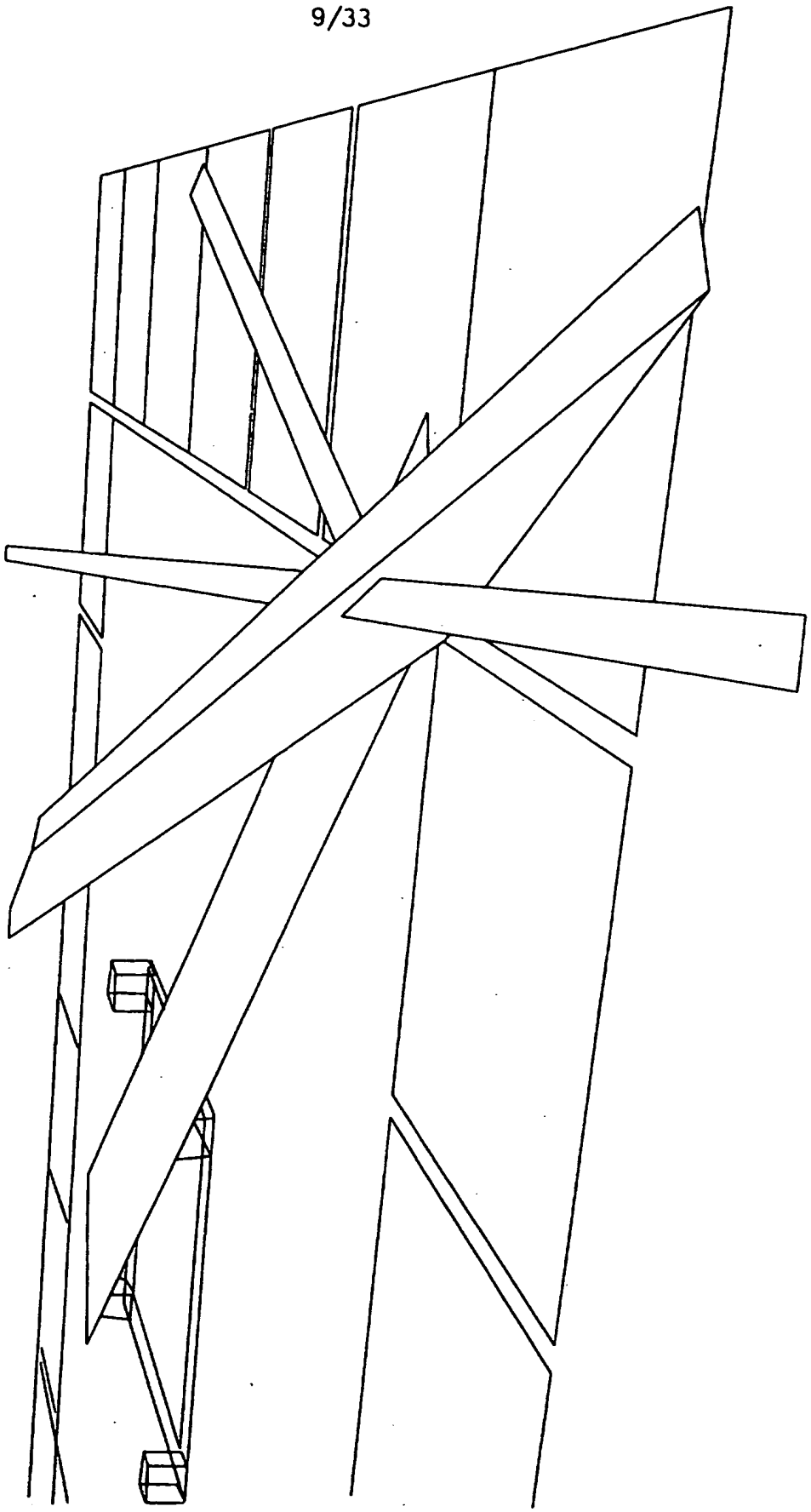


FIG. 13

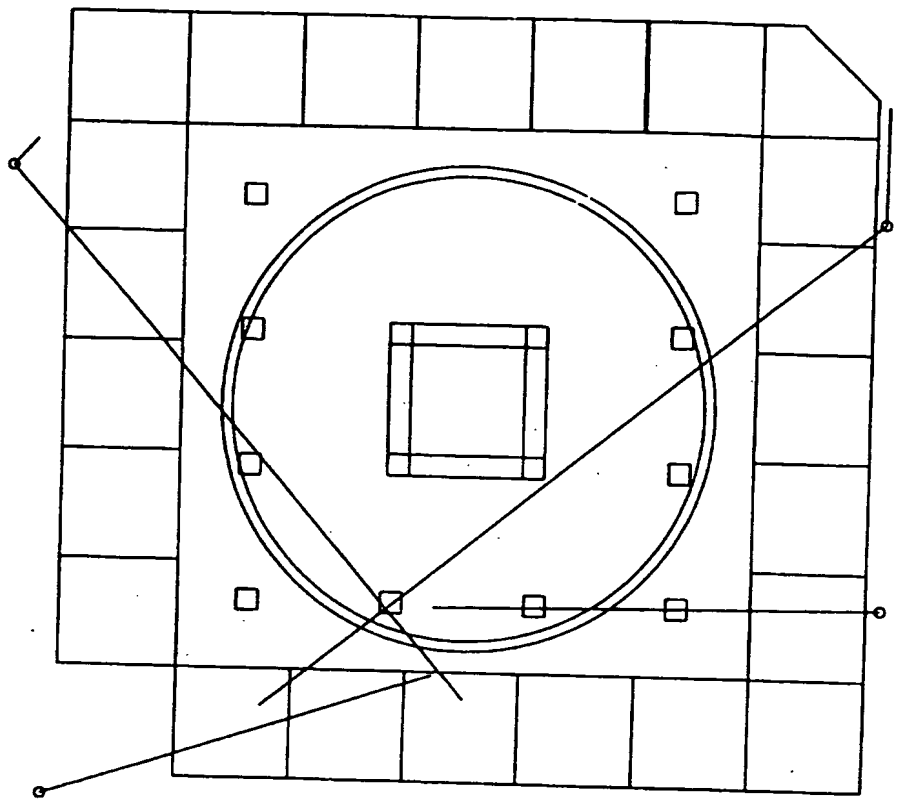


FIG. 14a

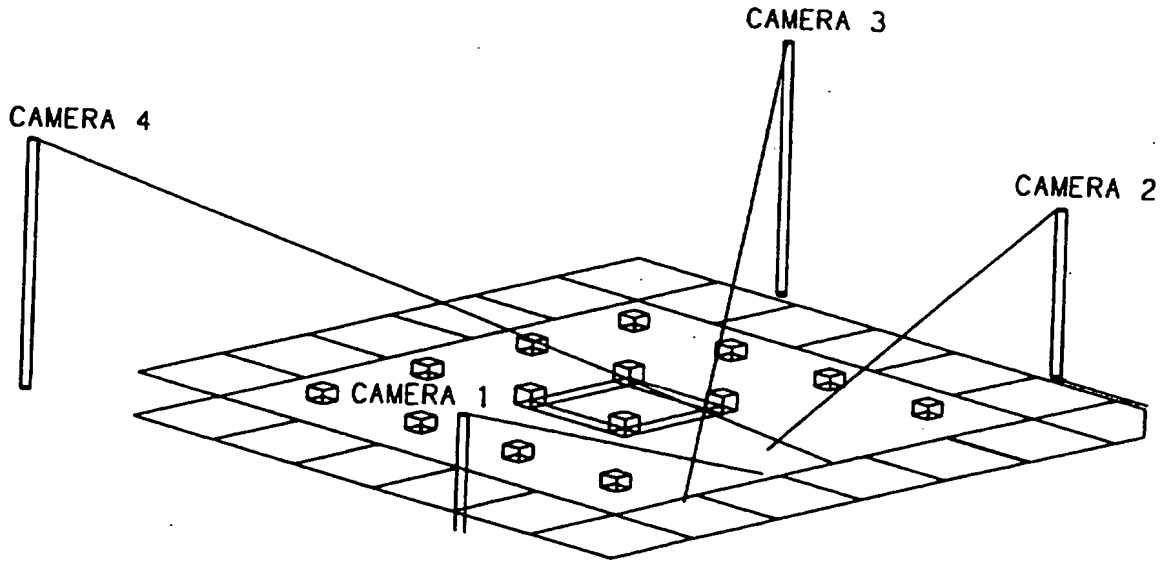


FIG. 14b

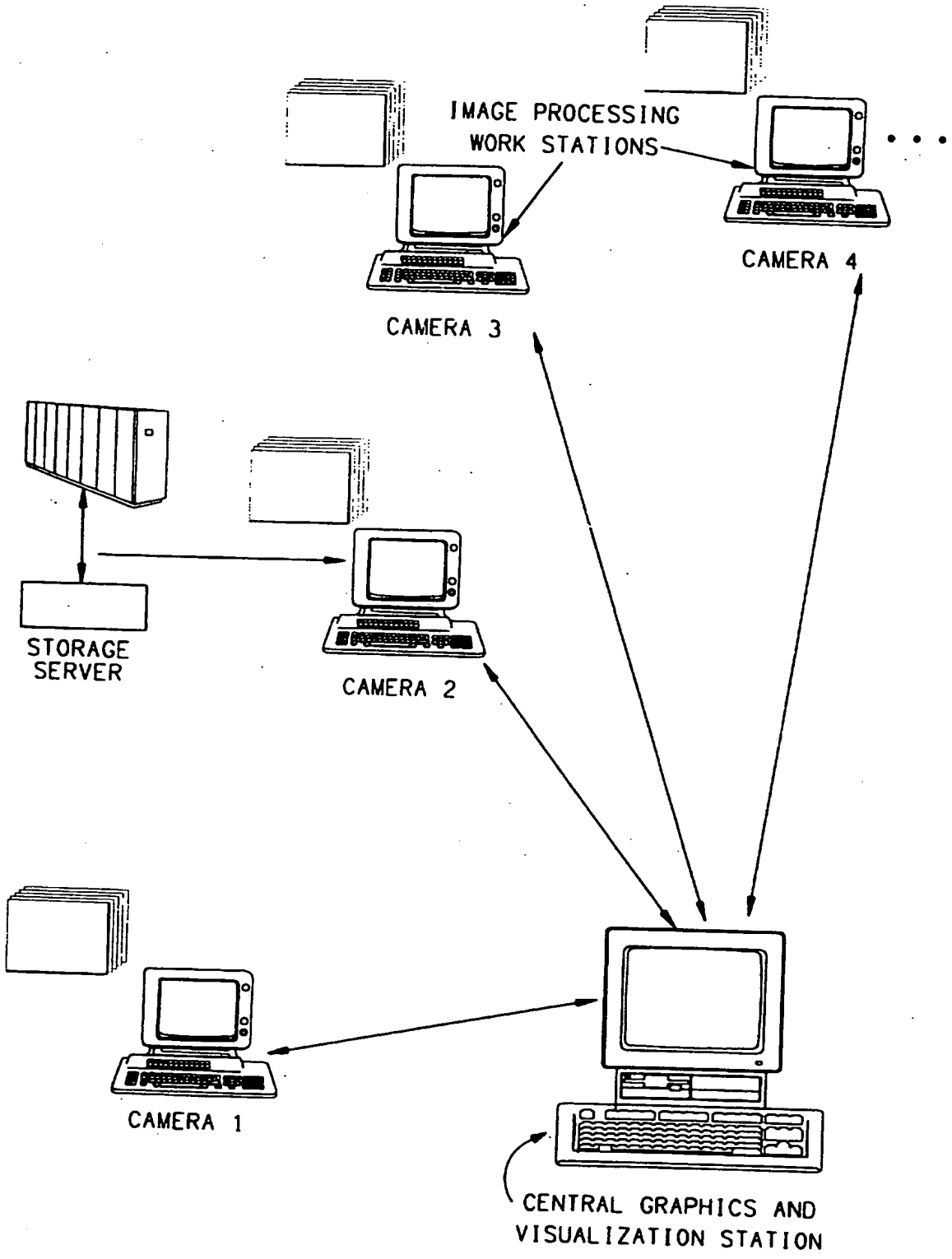


IMAGE PROCESSING
WORK STATIONS
CAMERA 4
CAMERA 3
STORAGE
SERVER
CAMERA 2
CAMERA 1
CENTRAL GRAPHICS AND
VISUALIZATION STATION

FIG. 15

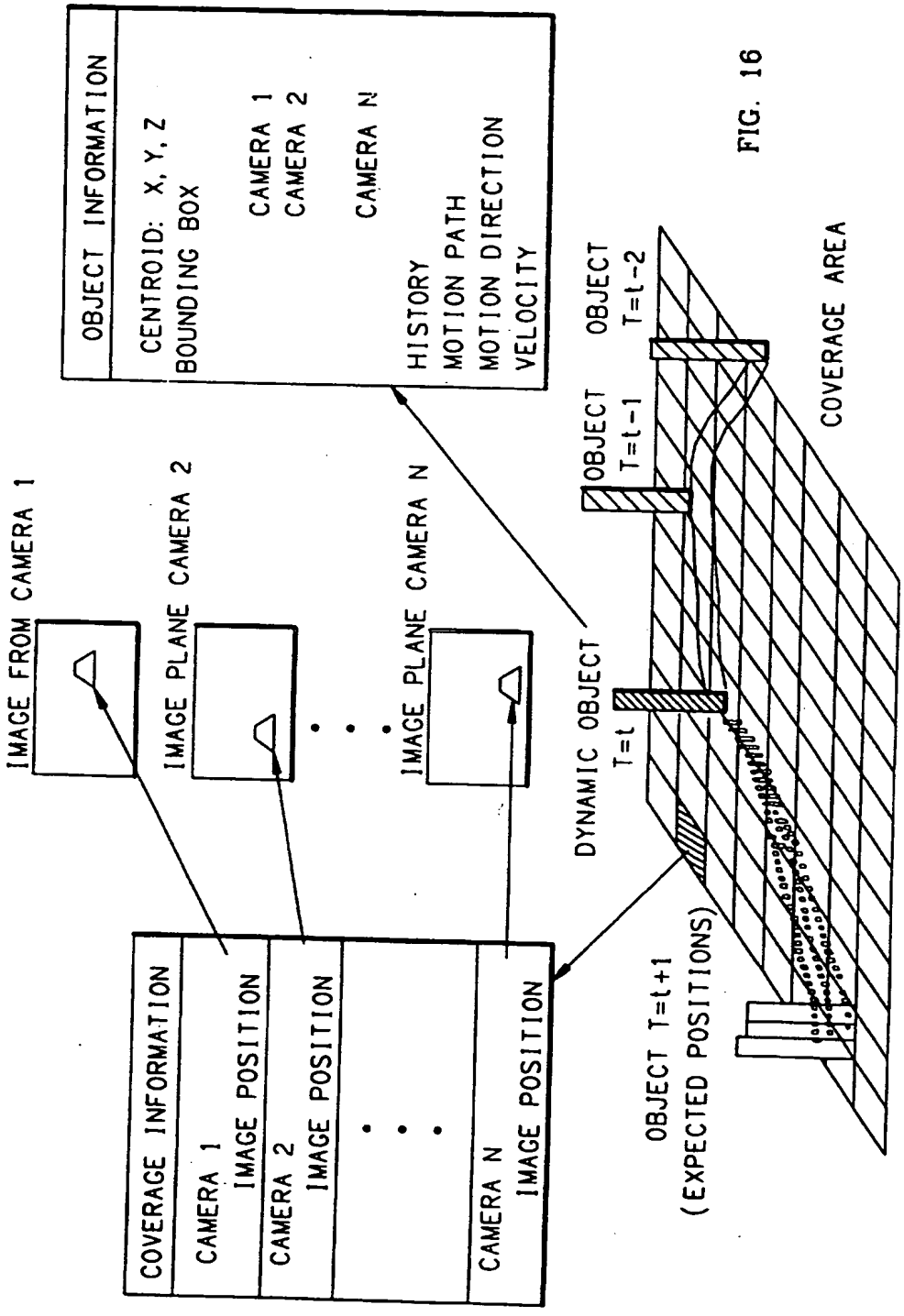
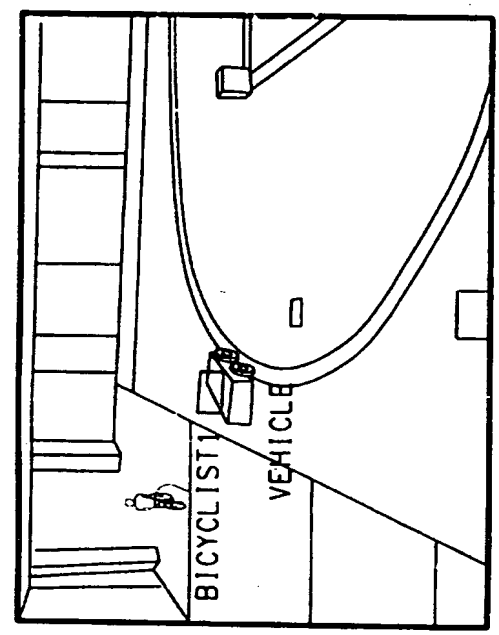
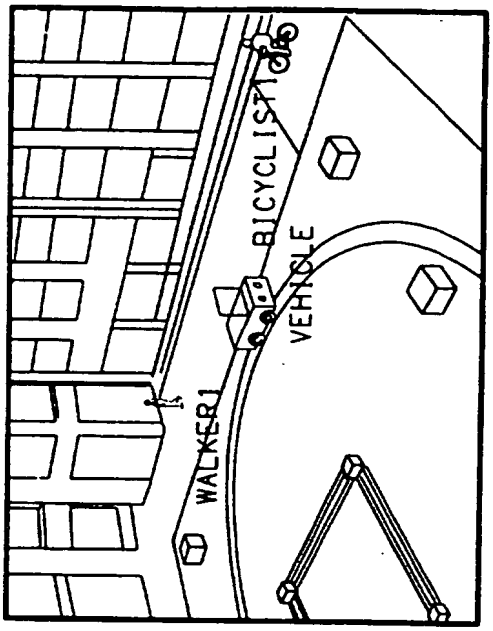

OBJECT INFORMATION
CENTROID: X, Y, Z
BOUNDING BOX
CAMERA 1
CAMERA 2
CAMERA N
HISTORY
MOTION PATH
MOTION DIRECTION
VELOCITY
IMAGE FROM CAMERA 1
IMAGE PLANE CAMERA 2
IMAGE PLANE CAMERA N
DYNAMIC OBJECT
T=t
OBJECT
T=t-1
OBJECT
T=t-2
COVERAGE AREA
OBJECT T=t+1
(EXPECTED POSITIONS)
COVERAGE INFORMATION
CAMERA 1
IMAGE POSITION
CAMERA 2
IMAGE POSITION
CAMERA N
IMAGE POSITION

FIG. 16



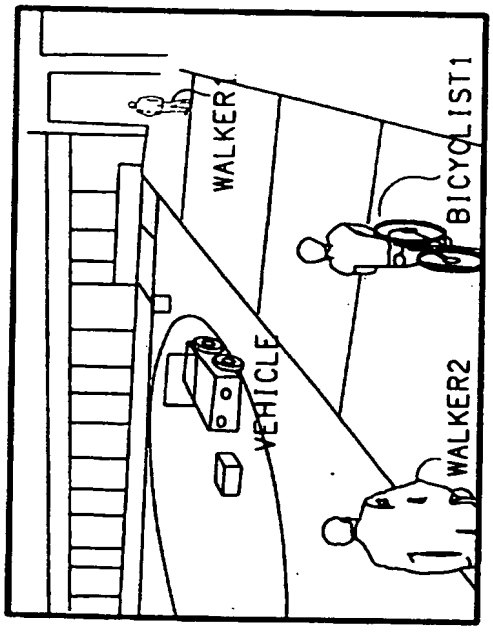
CAMERA 2
VIDEO FRAME 00: 21: 29: 15

FIG. 17b



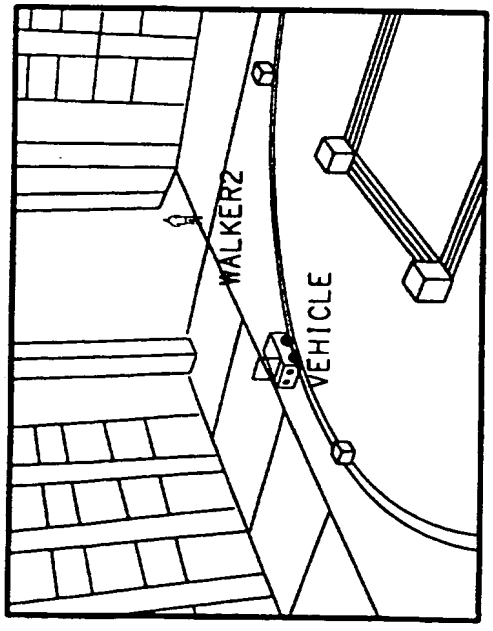
CAMERA 4
VIDEO FRAME 00: 20: 09: 18

FIG. 17d



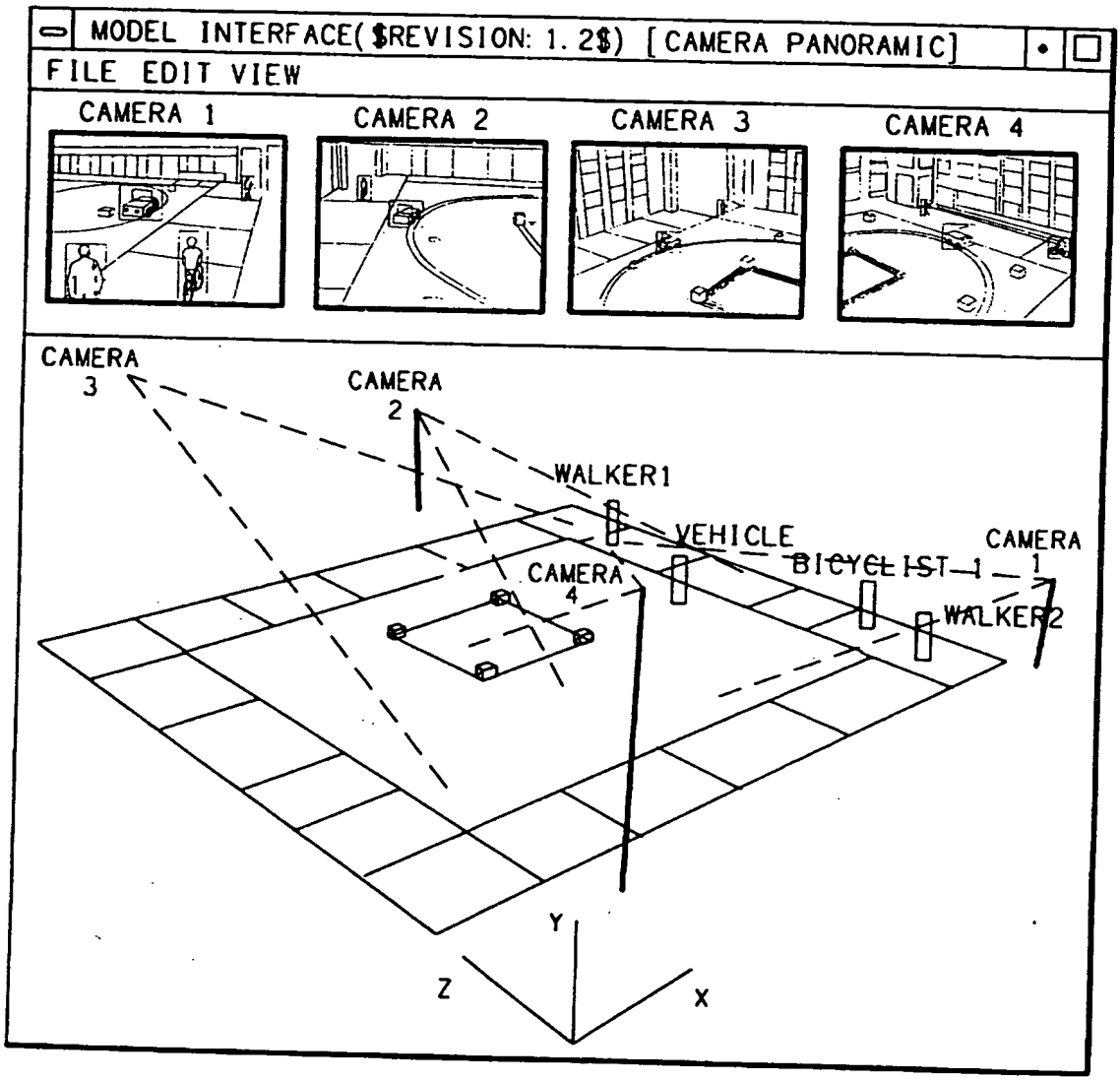
CAMERA 1
VIDEO FRAME 00: 21: 31: 24

FIG. 17a



CAMERA 3
VIDEO FRAME 00: 22: 29: 06

FIG. 17c

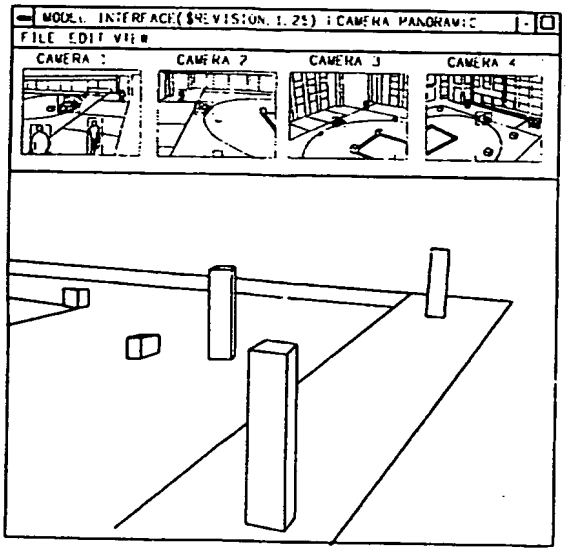


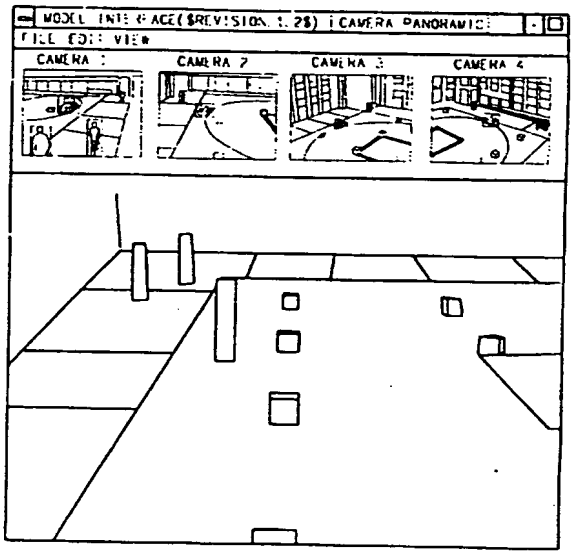
GLOBAL TIME 00: 22: 29: 06

FIG. 18

FIG. 19a

FIG. 19b





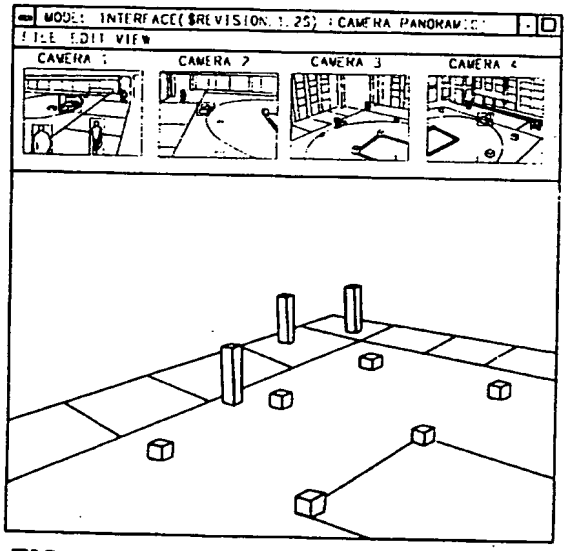


FIG. 19c

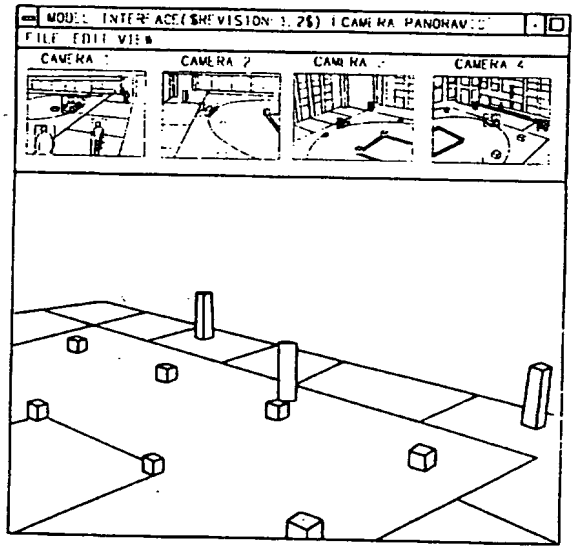


FIG. 19d

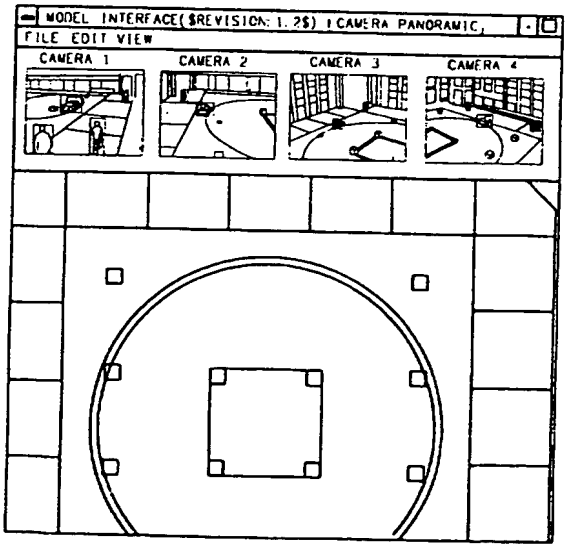
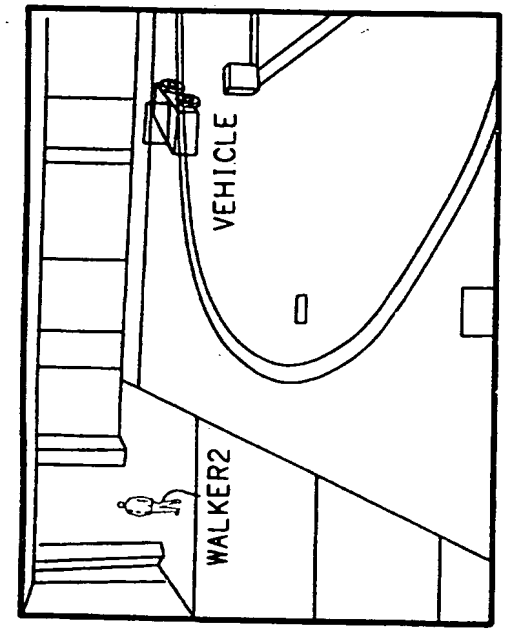
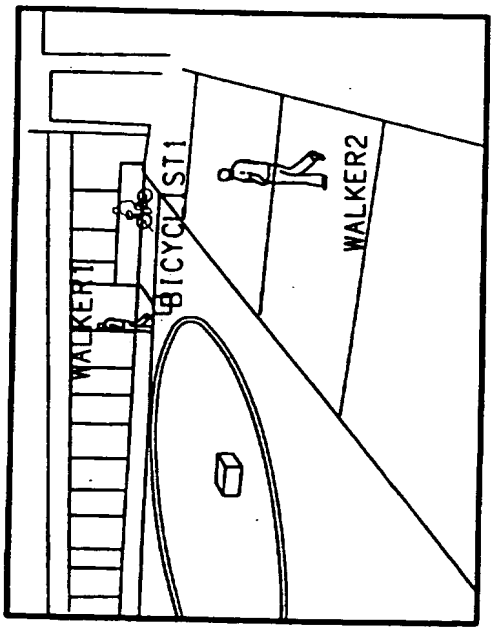


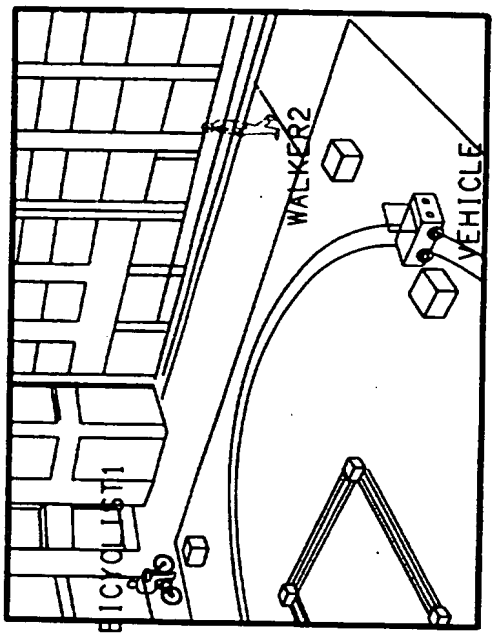
FIG. 19e



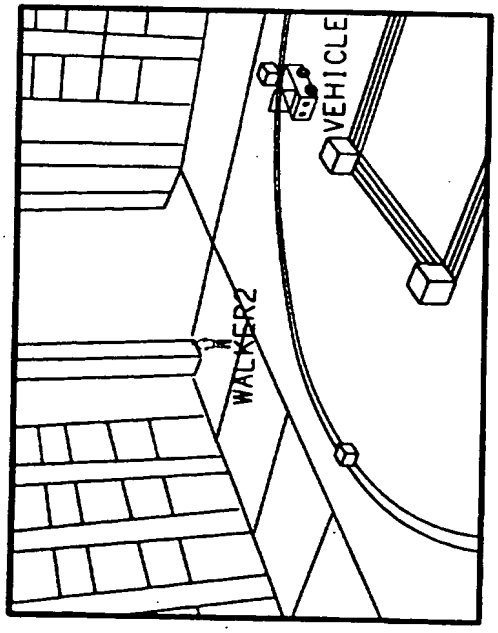
CAMERA 2
VIDEO FRAME 00: 21: 39: 15 FIG. 20b



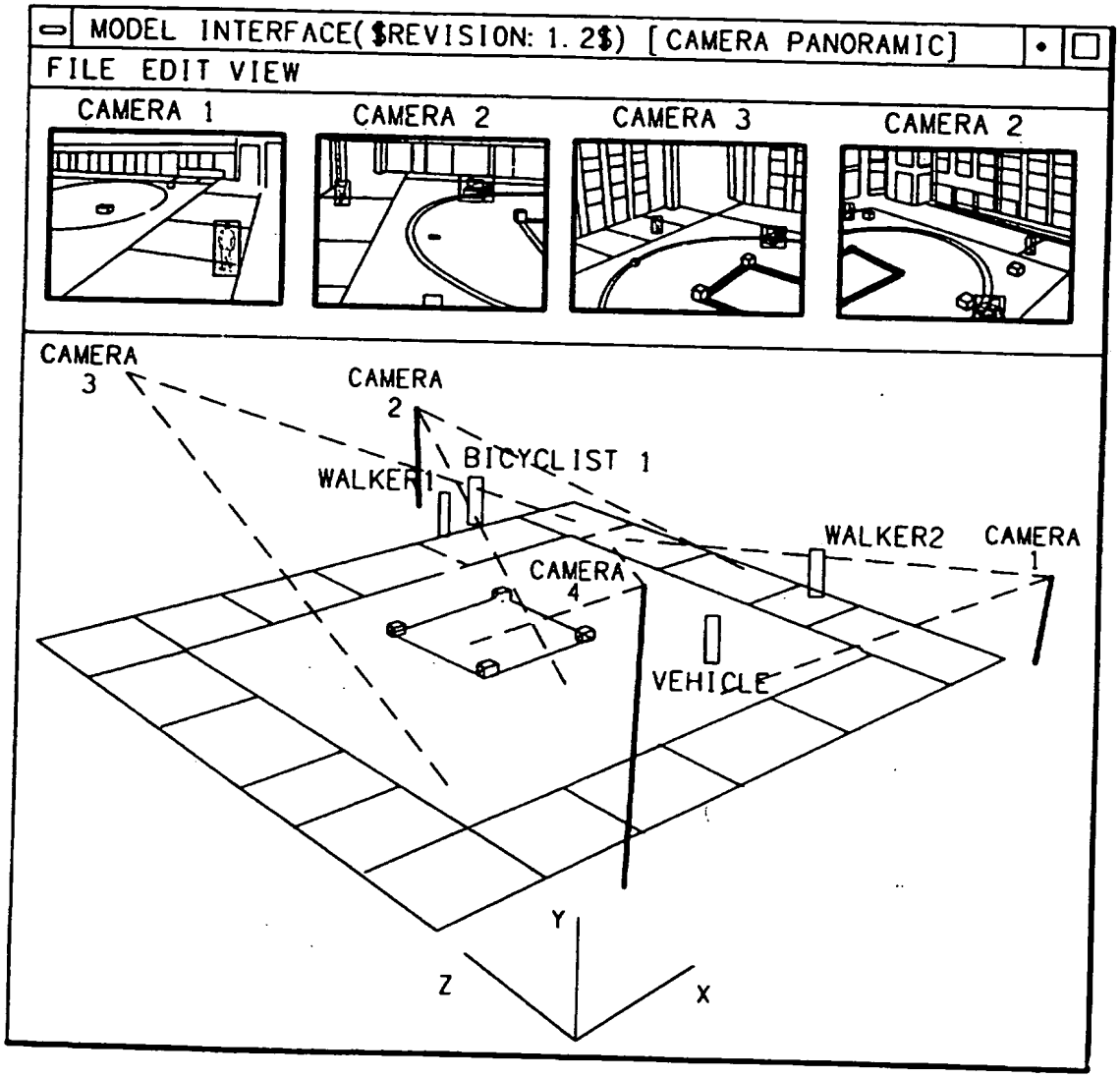
CAMERA 1
VIDEO FRAME 00: 21: 41: 24 FIG. 20a



CAMERA 4
VIDEO FRAME 00: 20: 19: 18 FIG. 20d



CAMERA 3
VIDEO FRAME 00: 22: 39: 06 FIG. 20c



GLOBAL TIME 00: 22: 39: 06

FIG. 21

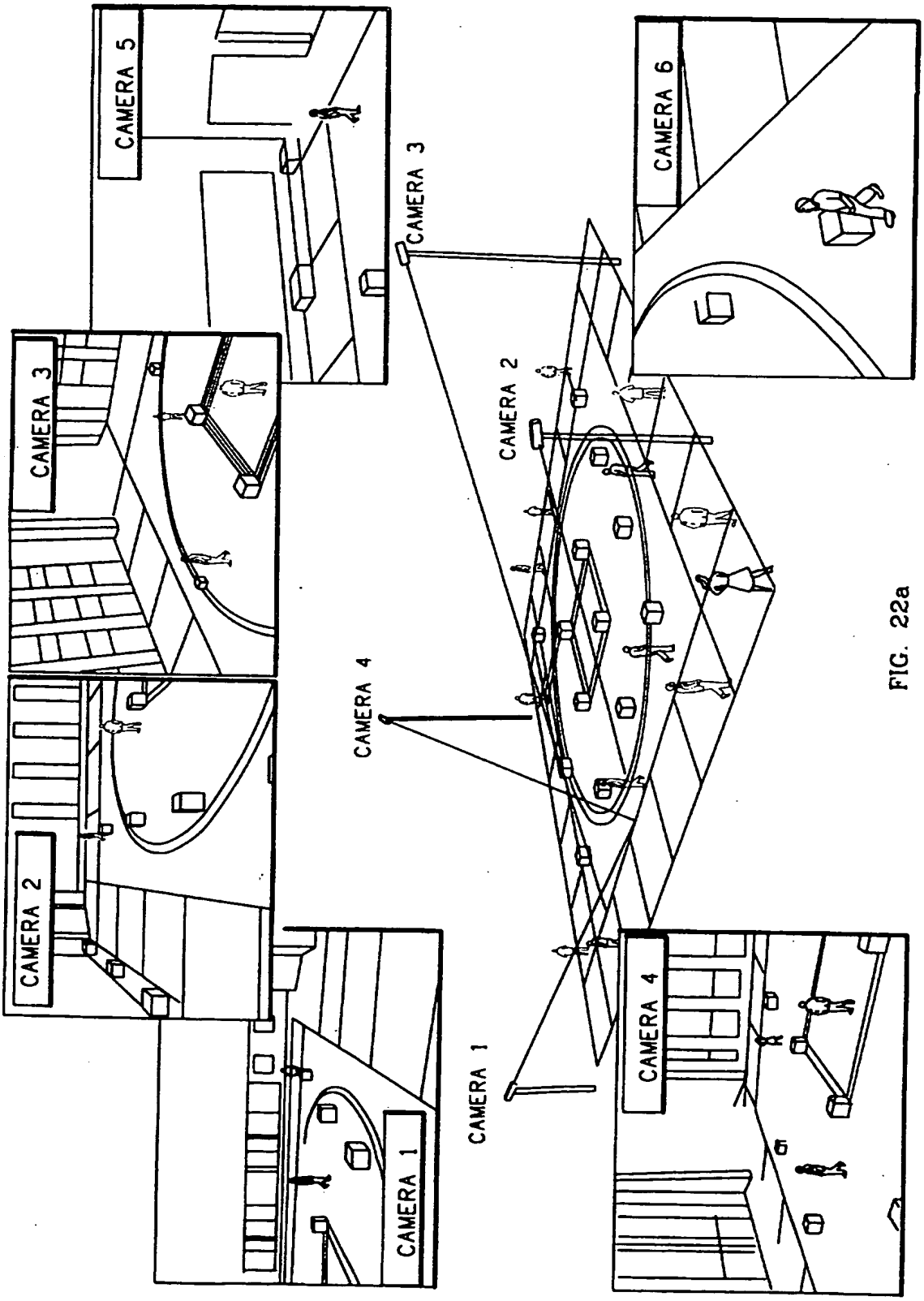

CAMERA 5
CAMERA 3
CAMERA 6
CAMERA 3
CAMERA 2
CAMERA 4
CAMERA 2
CAMERA 1
CAMERA 4
CAMERA 1

FIG. 22a

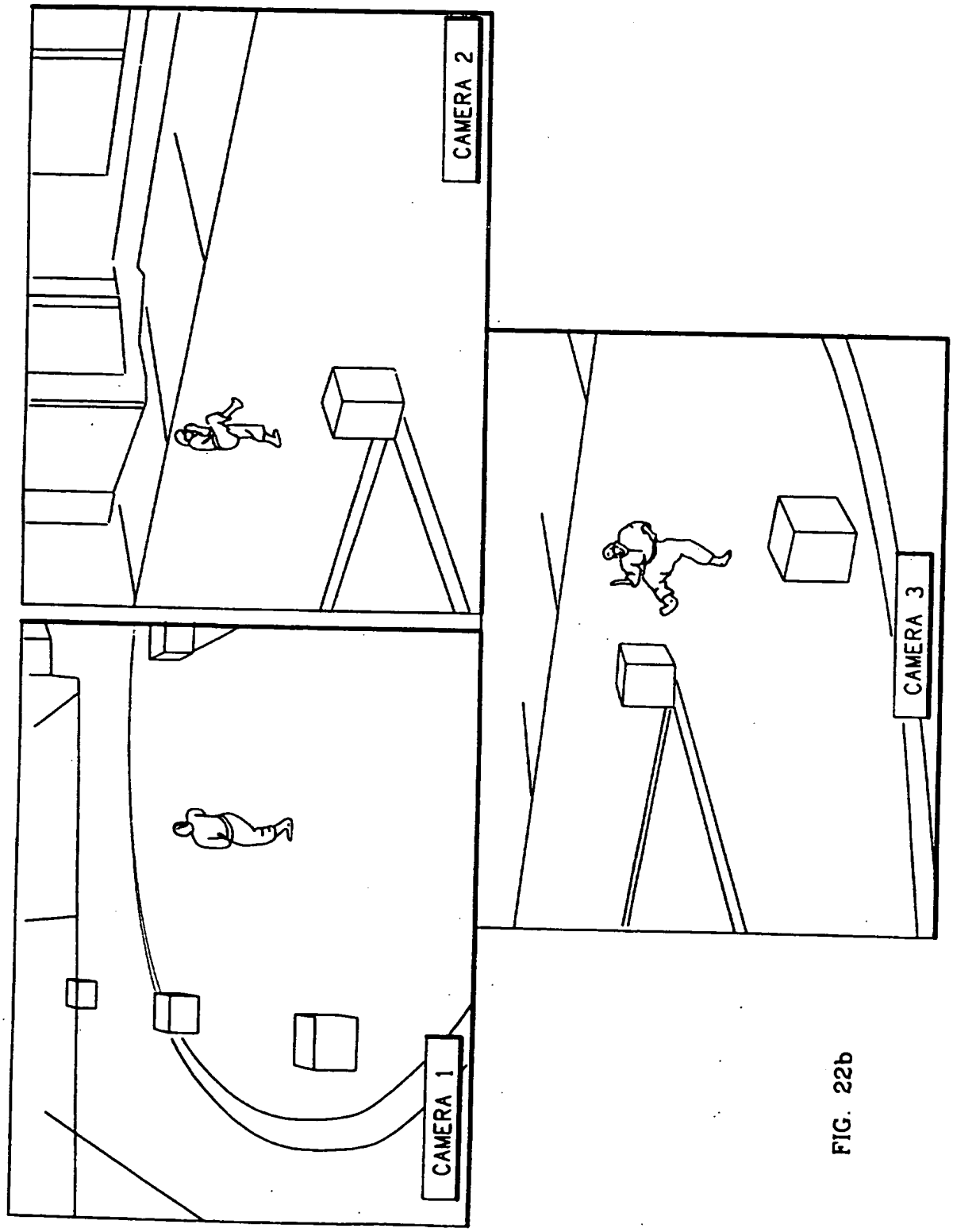


FIG. 22b

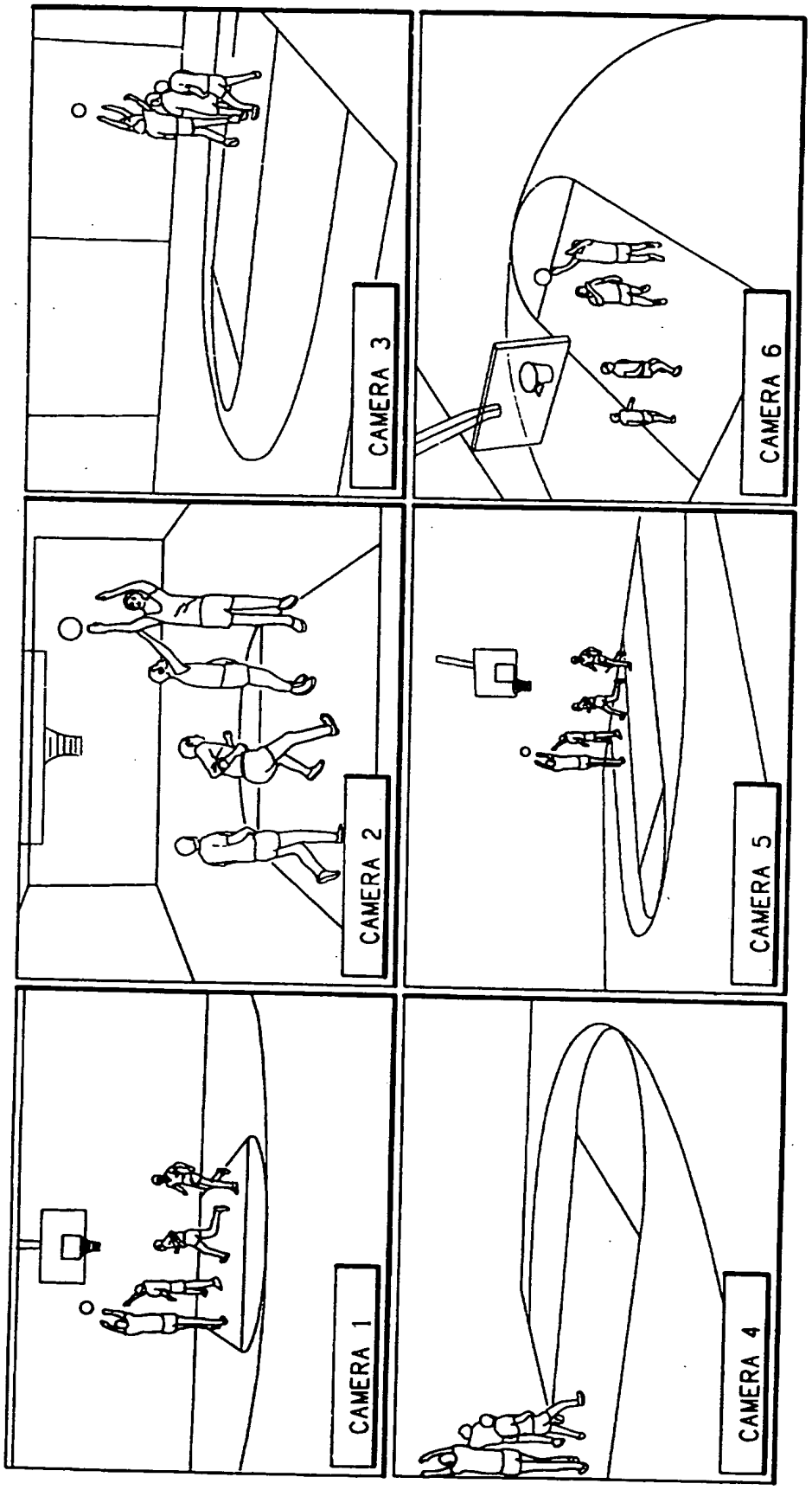


FIG. 22c

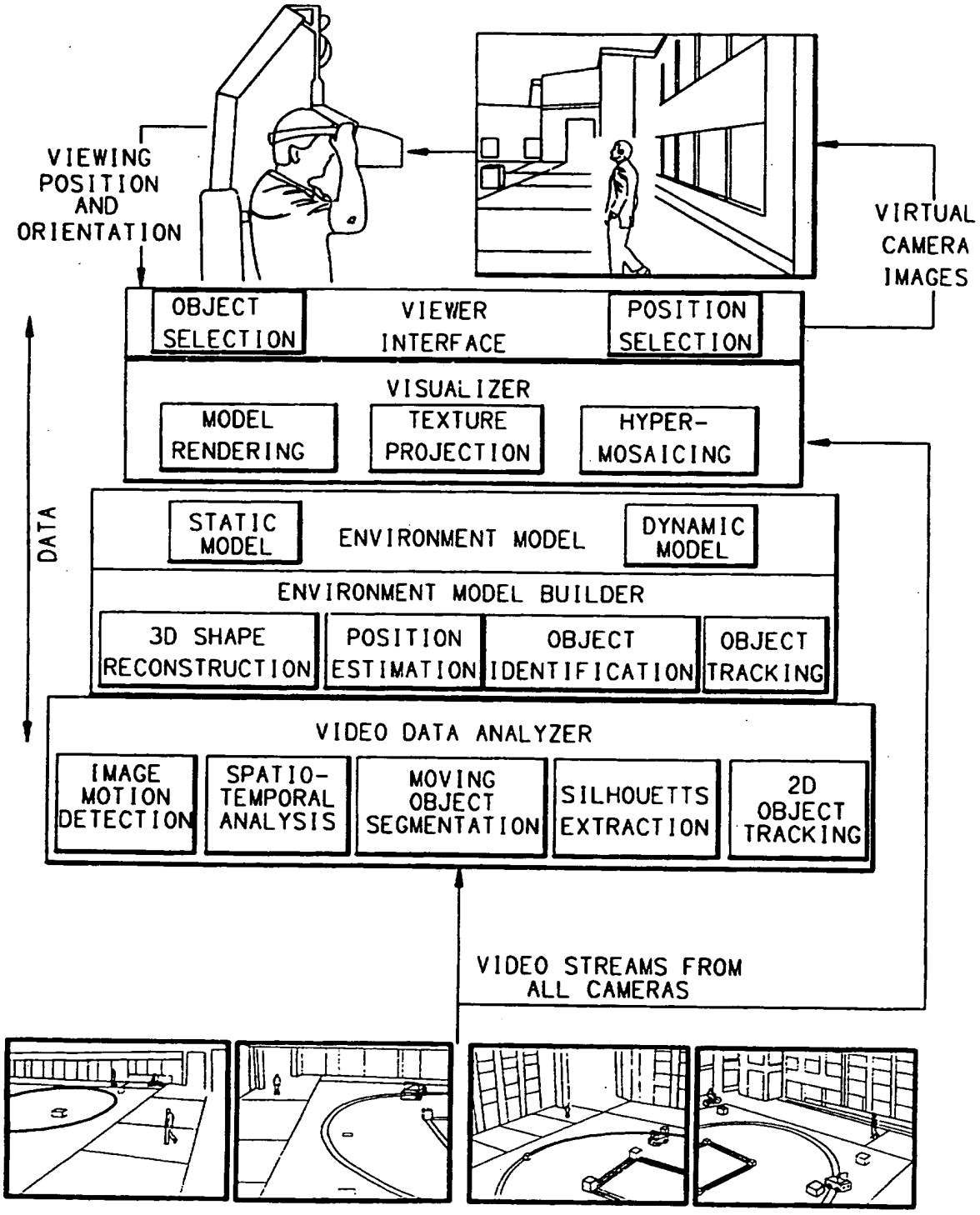

VIEWING POSITION AND ORIENTATION
VIRTUAL CAMERA IMAGES
OBJECT SELECTION
VIEWER INTERFACE
POSITION SELECTION
VISUALIZER
MODEL RENDERING
TEXTURE PROJECTION
HYPER-MOSAICING
DATA
STATIC MODEL
ENVIRONMENT MODEL
DYNAMIC MODEL
ENVIRONMENT MODEL BUILDER
3D SHAPE RECONSTRUCTION
POSITION ESTIMATION
OBJECT IDENTIFICATION
OBJECT TRACKING
VIDEO DATA ANALYZER
IMAGE MOTION DETECTION
SPATIO-TEMPORAL ANALYSIS
MOVING OBJECT SEGMENTATION
SILHOUETTS EXTRACTION
2D OBJECT TRACKING
VIDEO STREAMS FROM ALL CAMERAS

FIG. 23

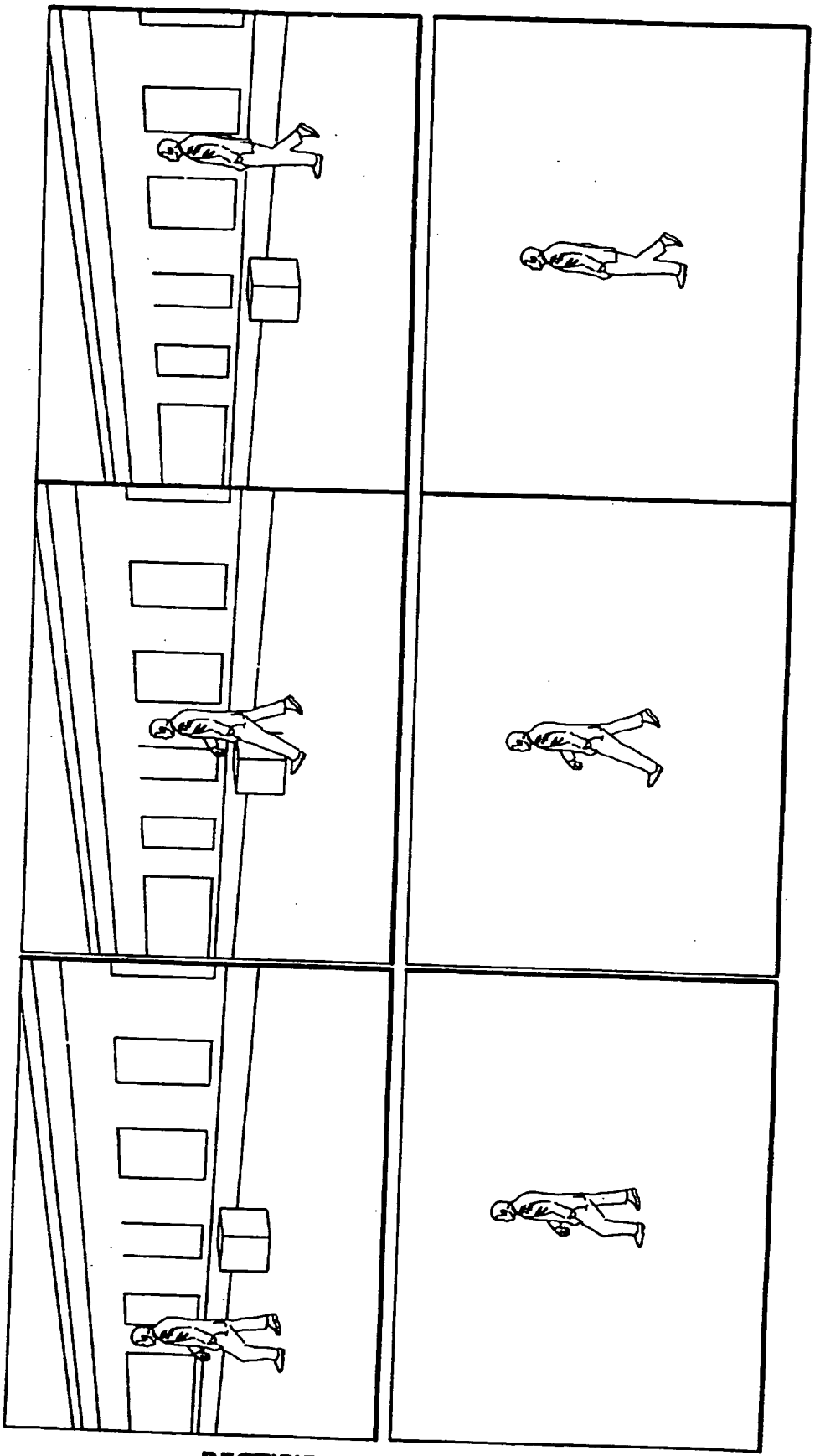


FIG. 24

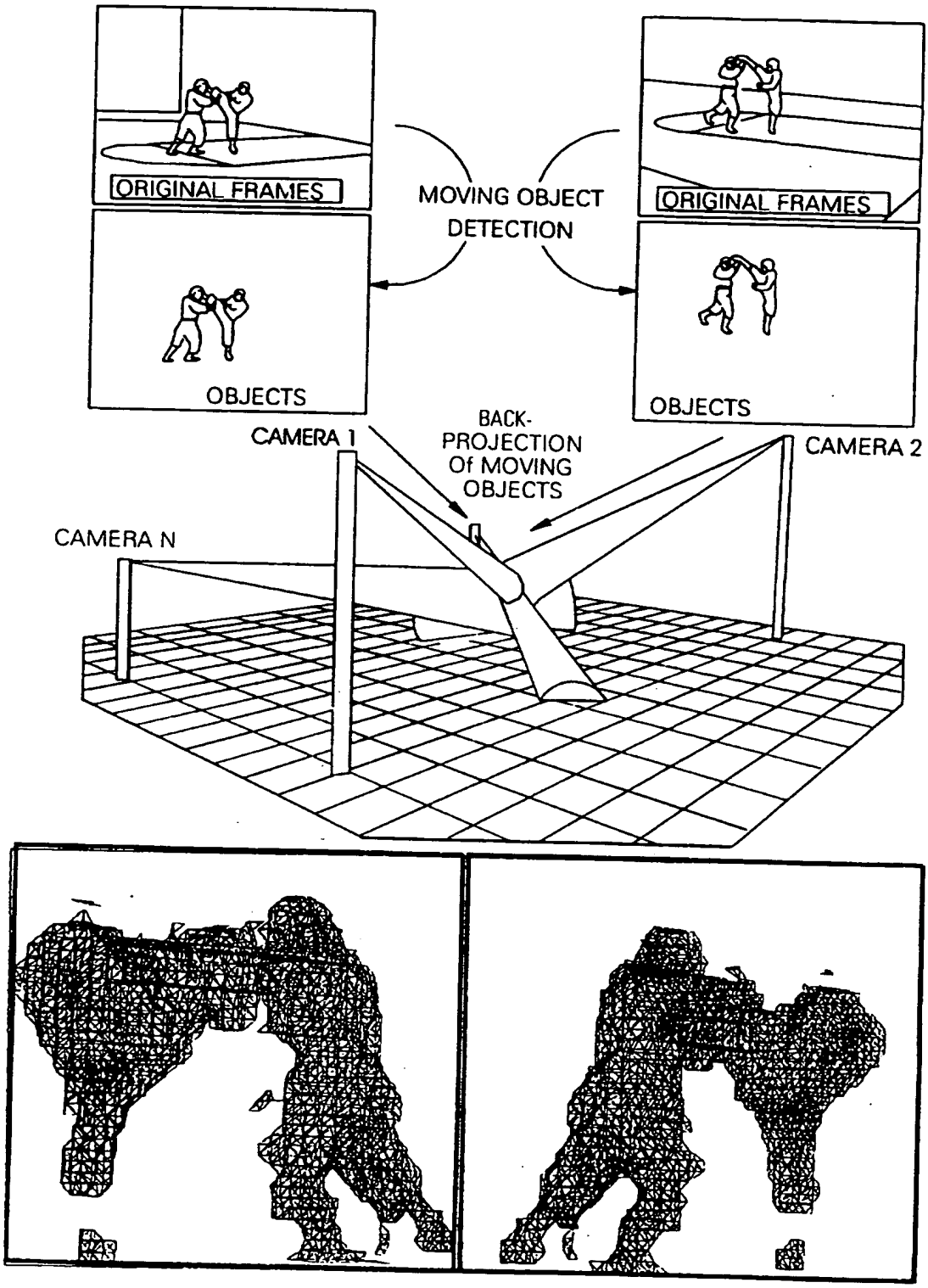

ORIGINAL FRAMES
MOVING OBJECT DETECTION
ORIGINAL FRAMES
OBJECTS
OBJECTS
CAMERA 1
BACK-PROJECTION Of MOVING OBJECTS
CAMERA 2
CAMERA N

FIG. 25

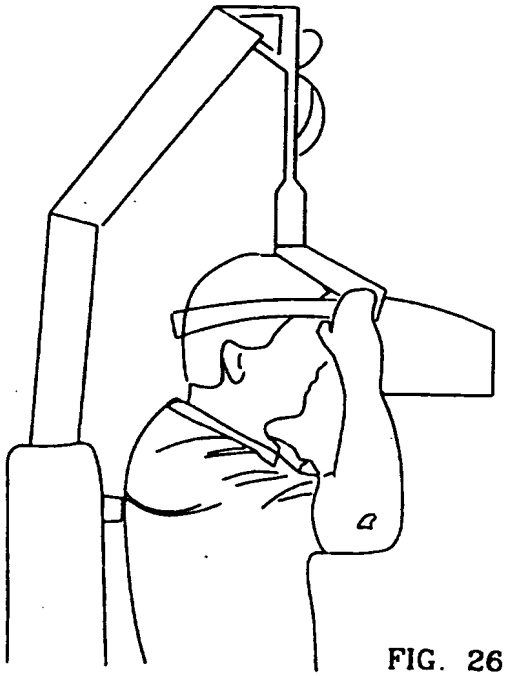


FIG. 26

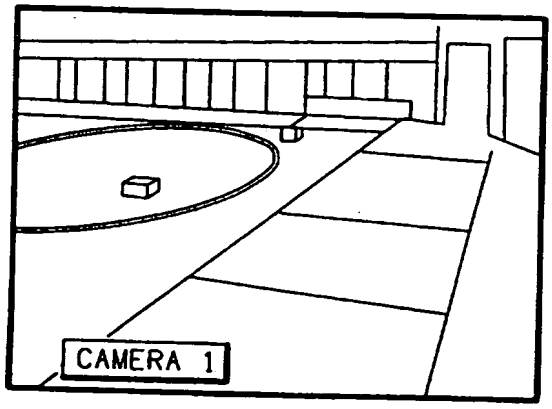


FIG. 27a

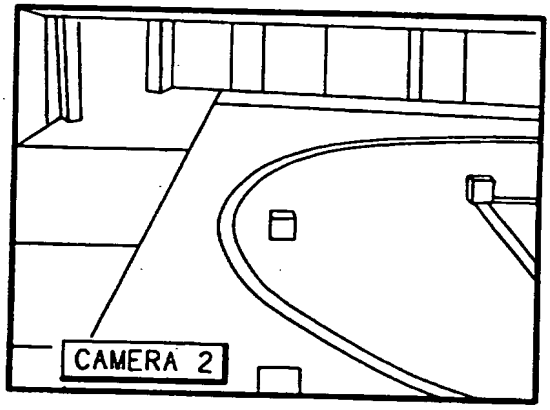


FIG. 27b

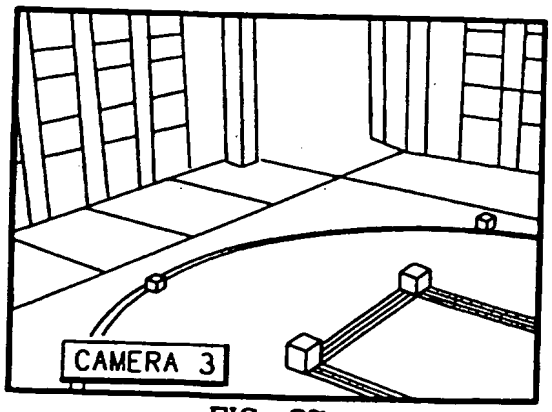


FIG. 27c

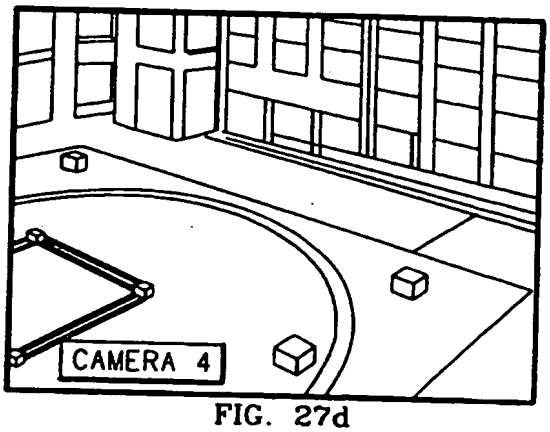


FIG. 27d

FIG. 28

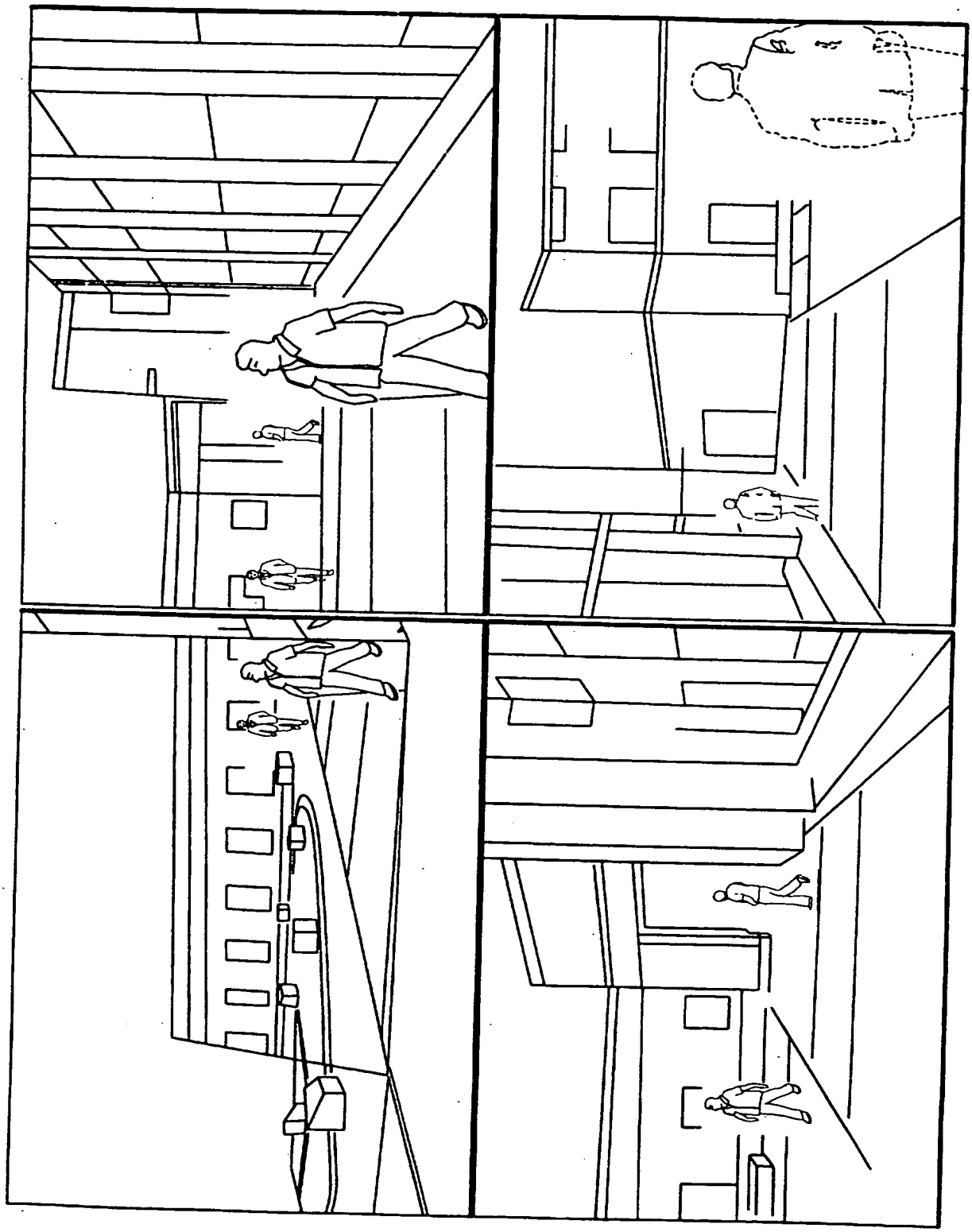
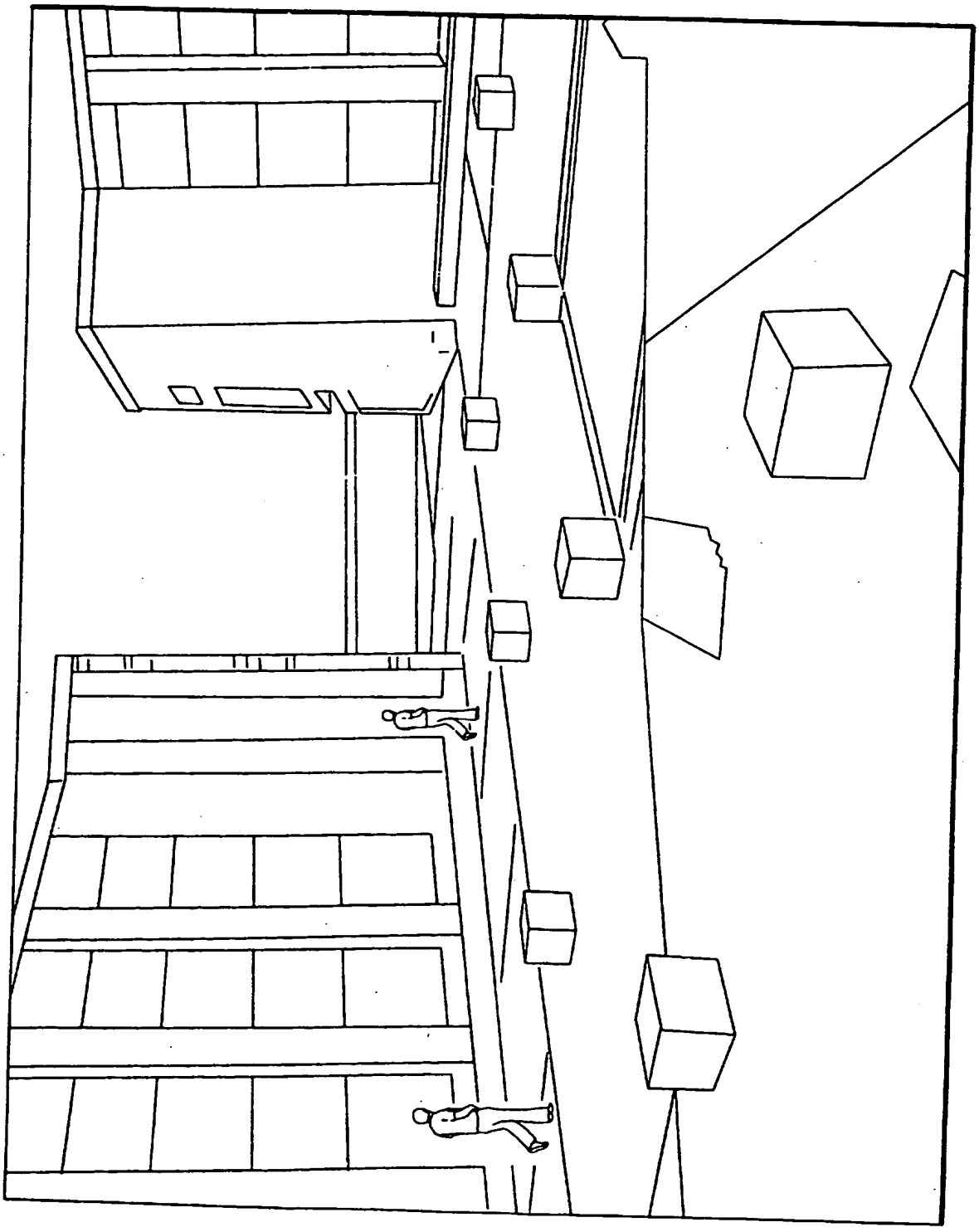


FIG. 29a



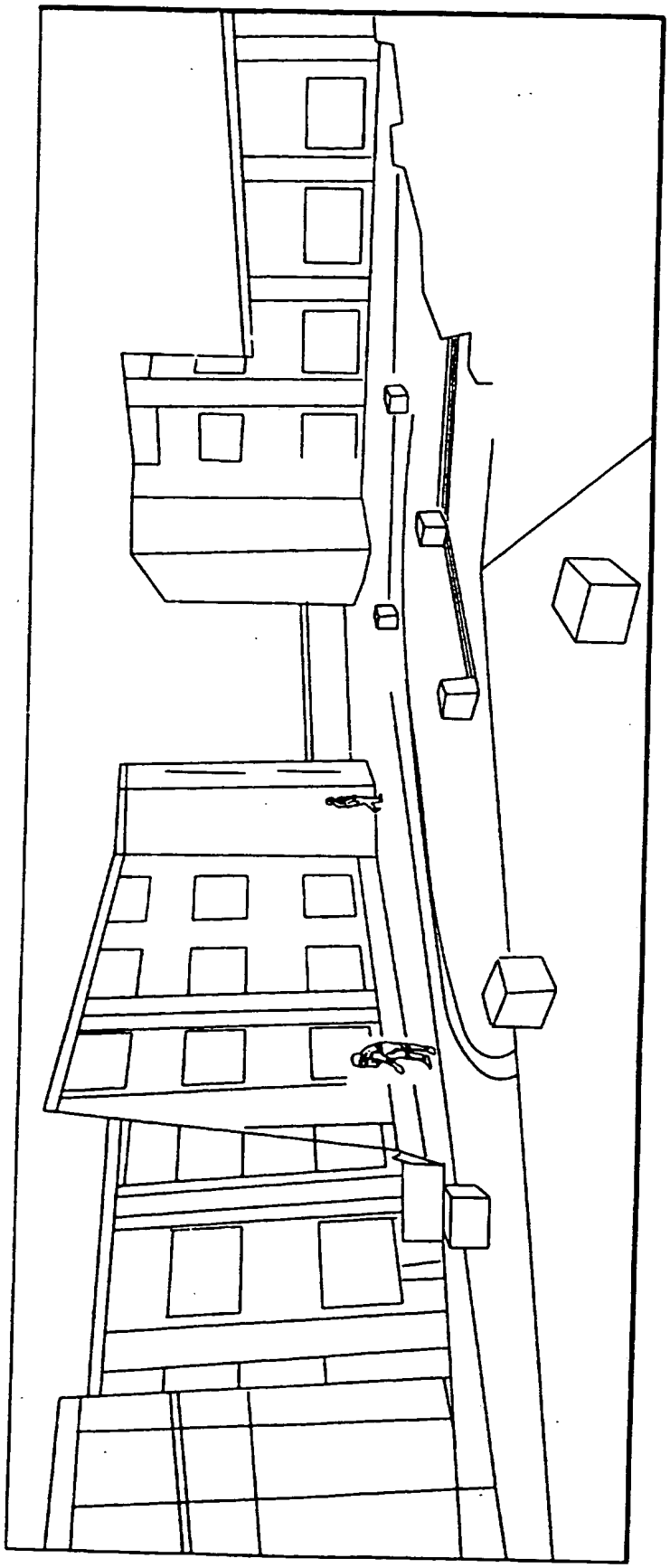


FIG. 29b

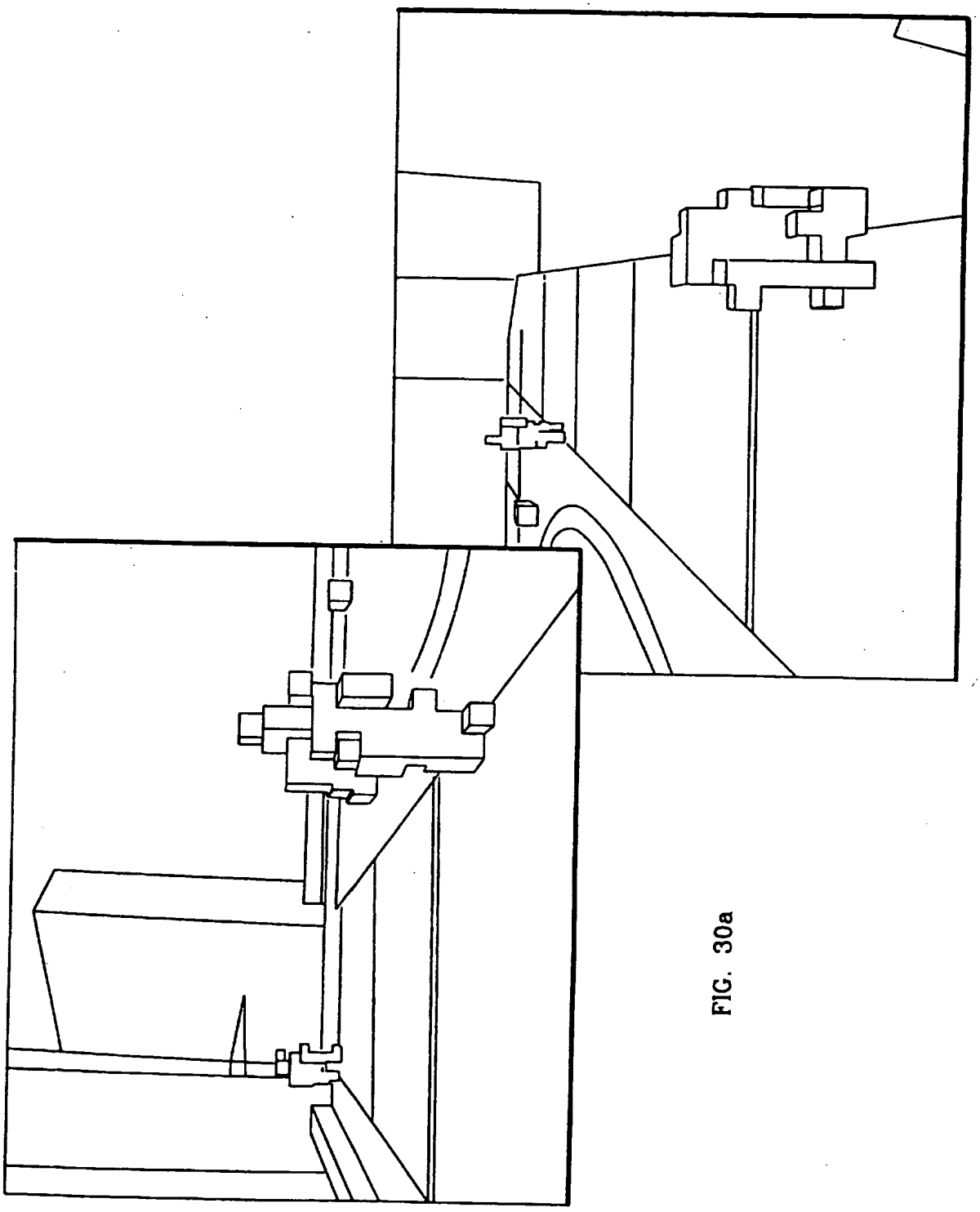


FIG. 30a

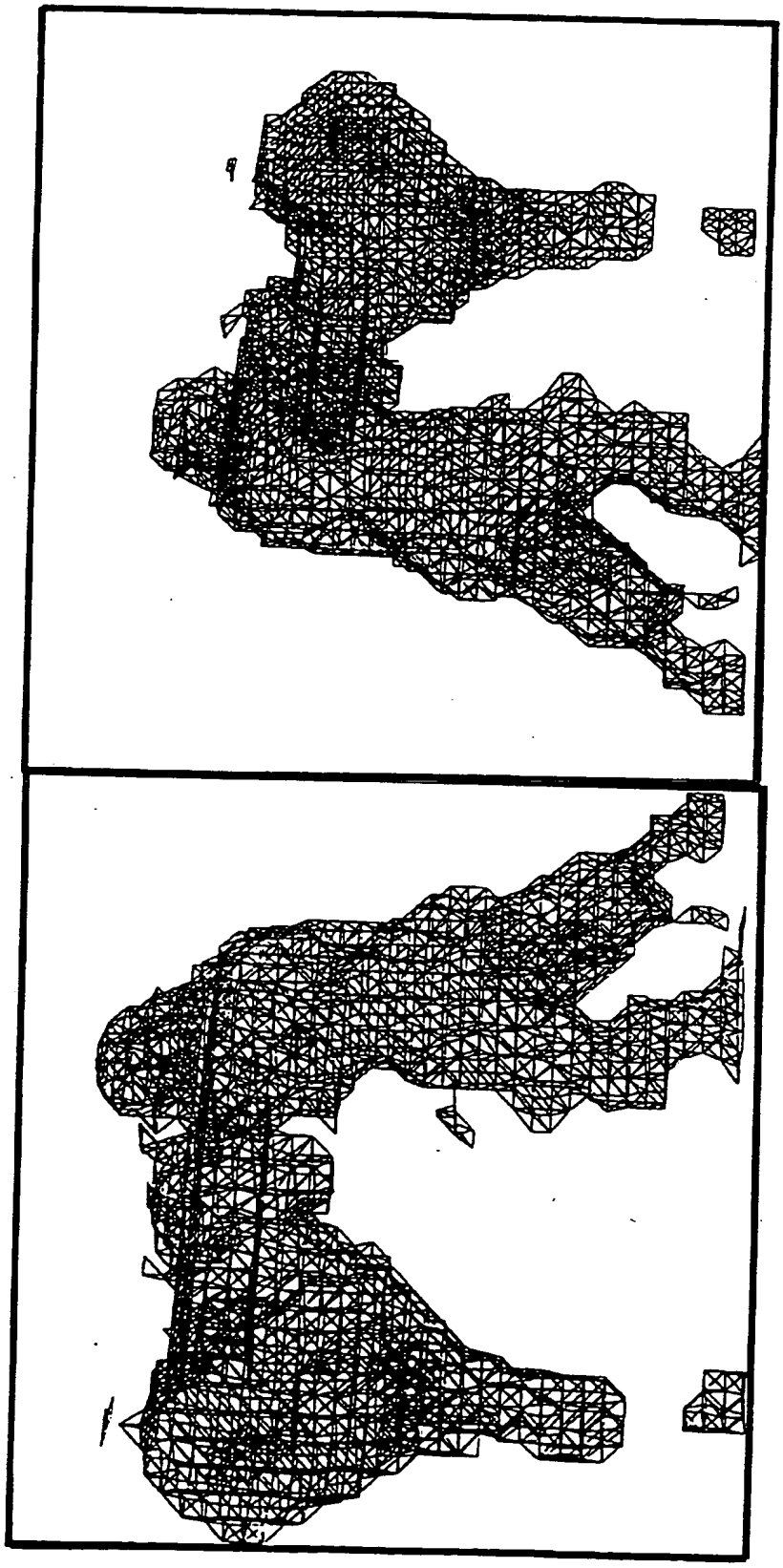


FIG. 30b

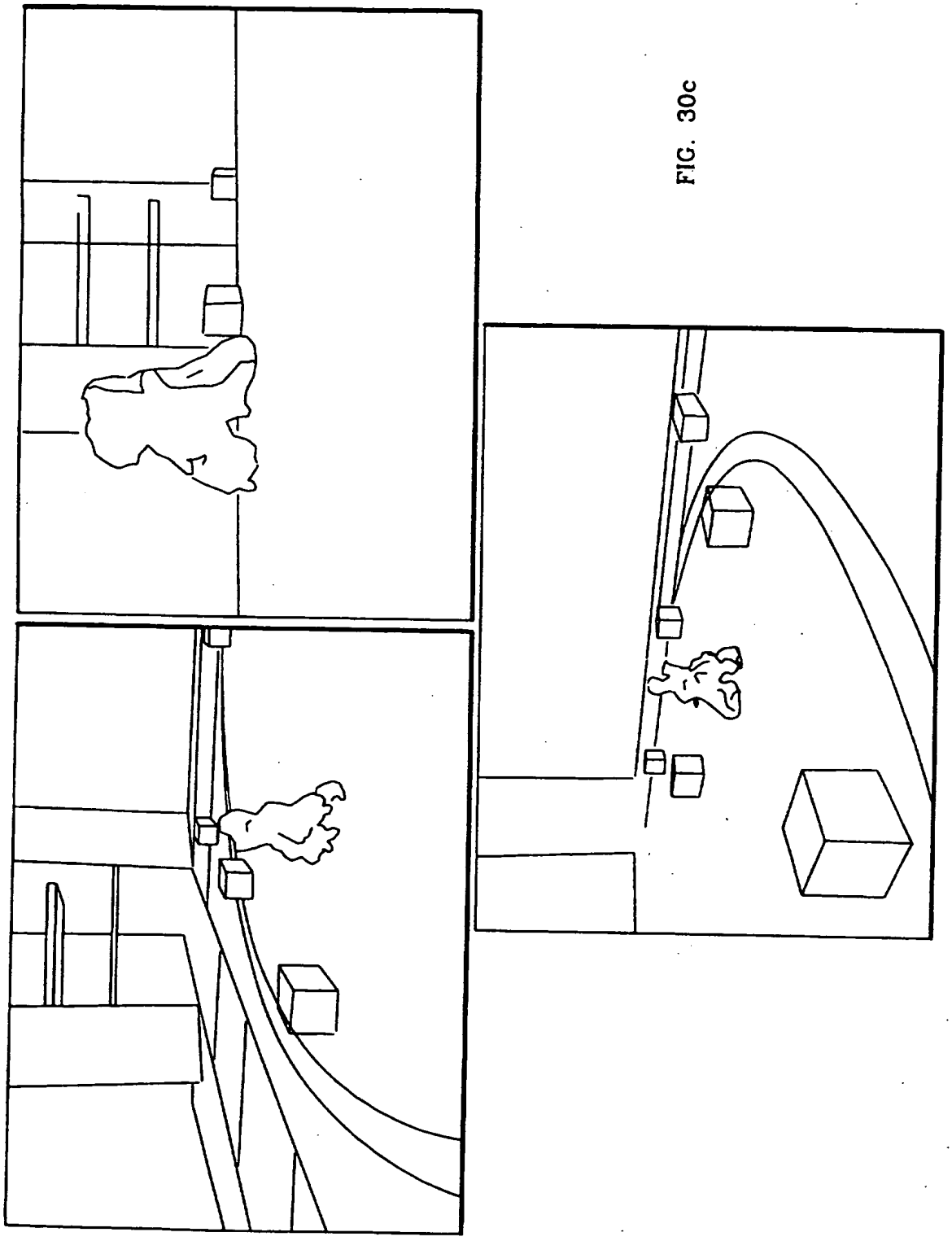
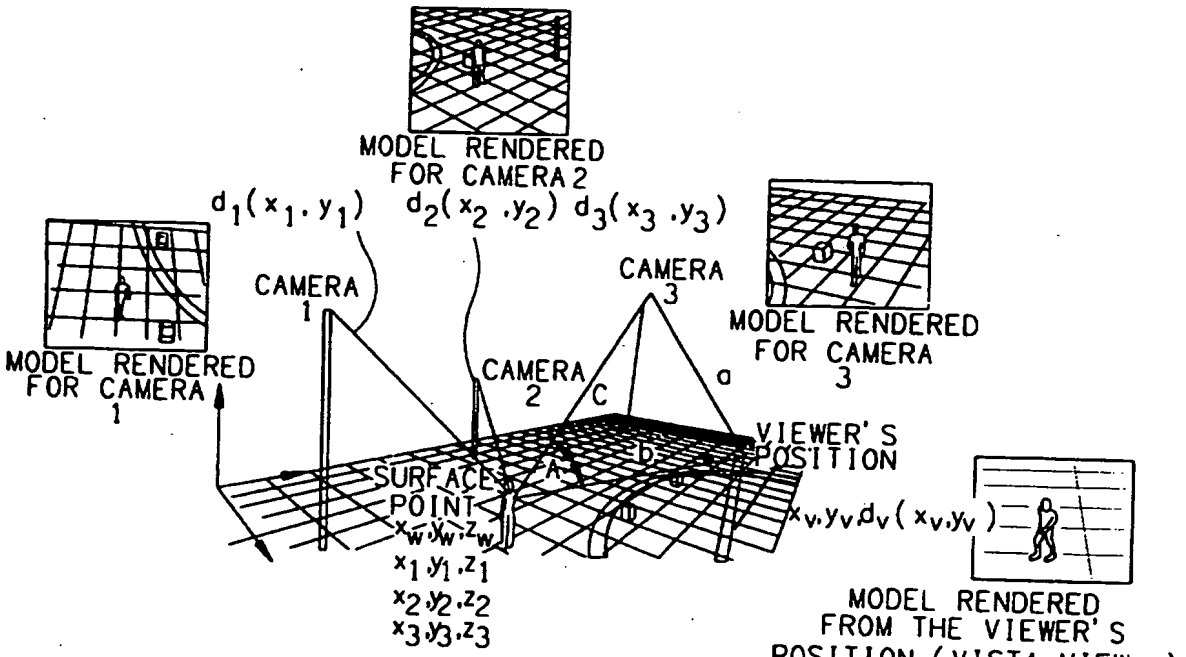


FIG. 30c




```
FOR EACH REAL CAMERA VIEW c do
  RENDER ENVIRONMENT MODEL FROM VIEW c;
  READ BACK DEPTH VALUES d_e(x,y)
END
FOR EACH VISTA VIEW v TO BE CREATED DO
  RENDER ENVIRONMENT MODEL FROM VISTA VIEW v;
  READ BACK DEPTH VALUES d_e(x,y)
  // CREATE VISTA VIEW BY FINDING INTENSITIES OF EACH PIXEL
    FOR EACH PIXEL (x_v, y_v) DO
      PROJECT (x_v, y_v, d_v(x_v, y_v)) ---> WORLD COORDINATES (x_w, y_w, z_w)
      FOR EACH REAL CAMERA VIEW c do
          PROJECT (x_w, y_w, z_w) --> CAMERA VIEW COORDINATES (x_c, y_c, z_c);
          // OCCLUSION TEST
        COMPARE d_c(x_c, y_c) AND z_c
      IF THEY ARE CLOSE ENOUGH THEN SIGNAL c AS A CANDIDATE VIEW_fi
      END
      FOR EACH CANDIDATE VIEW cv DO
COMPUTE EVALUATION CRITERION c_cv = f(arccos (b^2+c^2-a^2)/2bc, β·b_c(x_c, y_c))
WHERE a, b, c ARE AS SHOWN IN THE FIGURE 2, β IS A WEIGHTING CONSTANT.
        END
        INTENSITY F_v(x_v, z_v)(t) = F_w(x_bv, y_bv)(t)
        WHERE bv = ARG MIN_cv(c_v)
    END
END
```

FIG. 31

```
// CALCULATE VIEWING REGIONS CORRESPONDING TO A VOXEL FOR EACH CAMERA
FOR EACH VOXEL (X,Y,Z) DO
    COMPUTE 8 BOUNDING BOX POINTS FOR VOXEL (X,Y,Z)
    PROJECT EACH ONTO THE IMAGE PLANE OF EACH CAMERA VIEW
    DETERMINE MINy, MAXy, MINx, MAXx VIEWING REGIONS FOR EACH CAMERA
END
// DETERMINE IF VIEWING REGION CONTAINS MOVING OBJECTS
FOR EACH VOXEL (x,y,z) DO
    FOR EACH REAL CAMERA VIEW c DO
        // DETERMINE IF VIEWING REGION CONTAINS MOVING OBJECTS
        FOR ROW=MINy TO MAXy DO
            FOR COLUMN=MENx TO MAXx DO
                COMPUTE IF PIXELc(COLUMN, ROW) CONTAINS MOVING OBJECTS
                IF MOVING OBJECTS EXIST THEN INCREMENT VOXEL(x,y,z).VALUE FI
            END
        END
    END
// VISUALIZE VOXELS
FOR EACH VOXEL(x,y,z) DO
    IF VOXEL(x,y,z).VALUE > THRESHOLD THEN VISUALIZE VOXEL(x,y,z) FI
END
```

FIG. 32

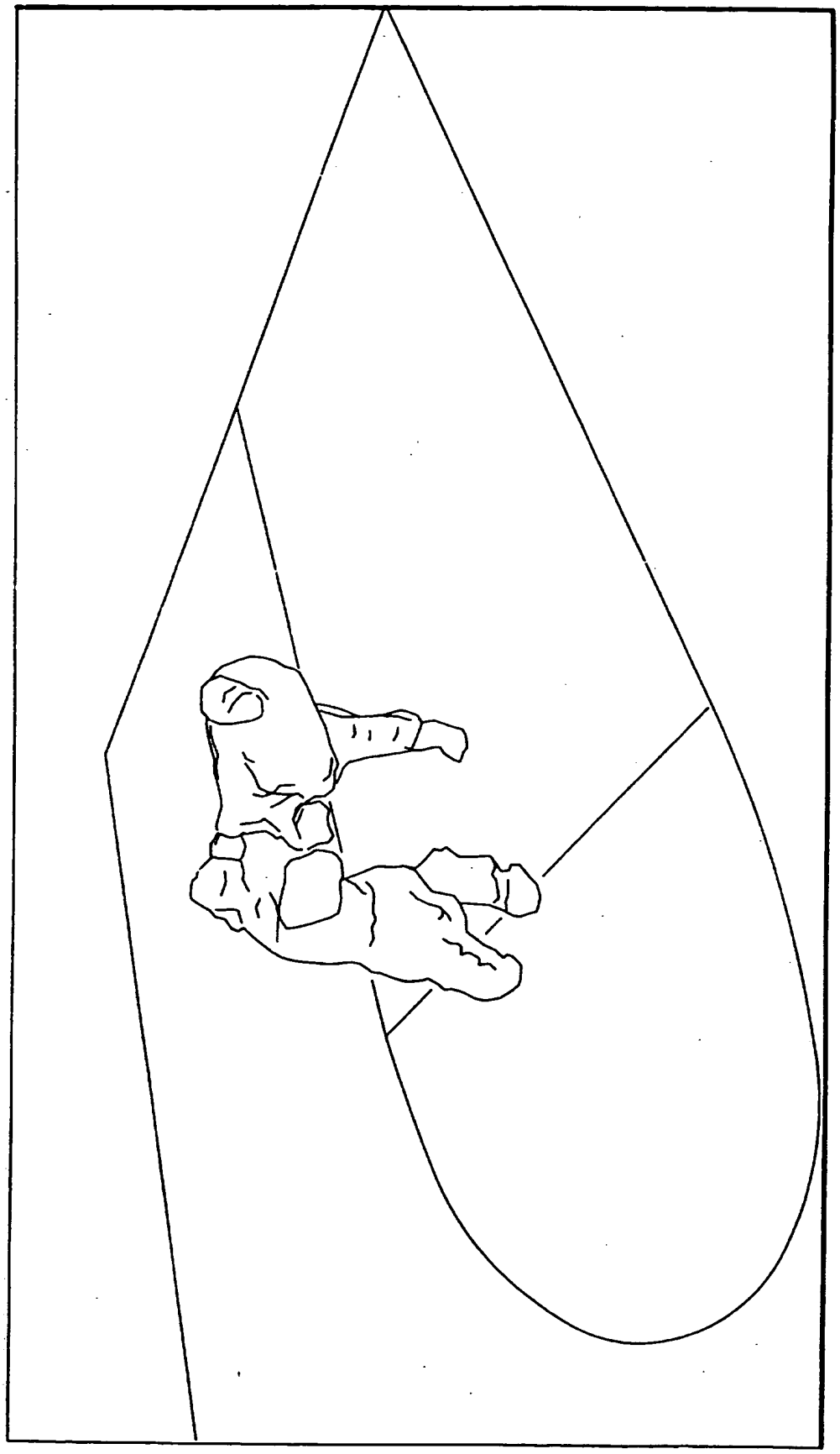


FIG. 33

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06K 09/20 — A3

(11) International Publication Number: WO 96/31047

(43) International Publication Date: 3 October 1996 (03.10.96)

(21) International Application Number: PCT/US96/04400

(22) International Filing Date: 29 March 1996 (29.03.96)

(30) Priority Data:
08/414,437 31 March 1995 (31.03.95) US
08/554,848 7 November 1995 (07.11.95) US

(71) Applicant *(for all designated States except US)*: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA [US/US]; 22nd floor, 300 Lakeside Drive, Oakland, CA 94612-3550 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: JAIN, Ramesh [US/US]; 4715 Reedly Terrace, La Jolla, CA 92130 (US). WAKIMOTO, Koji [JP/US]; 6-36-8-204, Shonandai, Fujisawa, Kanagawa 252 (JP). MOEZZI, Saied [US/US]; 10420 Caminito Alvarez, San Diego, CA 92126 (US). KATKERE, Arun [IN/US]; 9500 Gilman Drive, La Jolla, CA 92093-0407 (US).

(74) Agent: FUESS, William, C.; Suite II-G, 10951 Sorrento Valley Road, San Diego, CA 92121-1613 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IS, JP, KE, KG, KP, KR, KZ, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, US, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

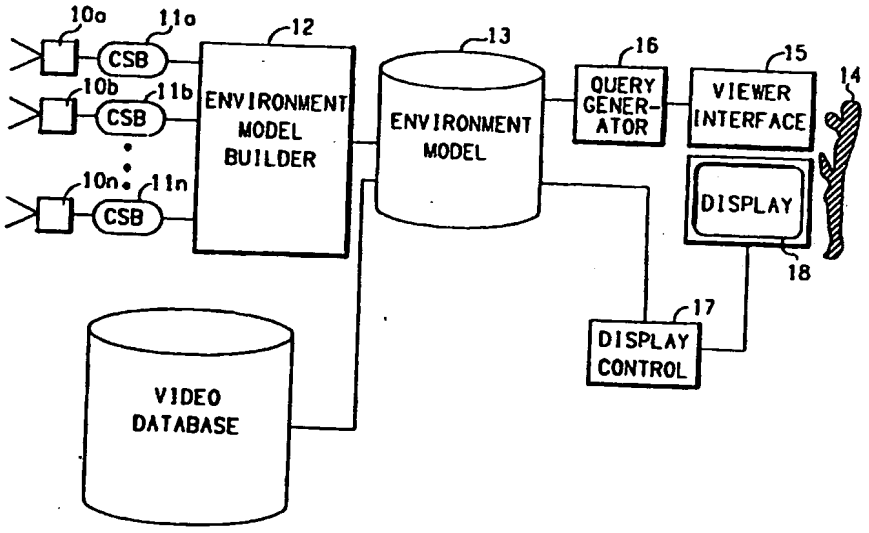
Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(88) Date of publication of the international search report: 7 November 1996 (07.11.96)

(54) Title: IMMERSIVE VIDEO



(57) Abstract

Immersive video, or television, images of a real-world scene are synthesized (i) on demand, (ii) in real time, (iii) as linked to any of a particular perspective on the scene, or an object or event in the scene, (iv) in accordance with user-specified parameters of presentation, including panoramic or magnified presentations, and/or (v) stereoscopically. Multiple video cameras (10a-10c) each at a different spatial location produce multiple two-dimensional video images of the scene. A viewer/user specifies viewing criterion (ia) at a viewer interface (10). A video display (18) receives and displays the synthesized 2-D video image(s).

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia
AT	Austria
AU	Australia
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CS	Czechoslovakia
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgystan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakhstan
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/04400

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G06K 09/20
US CL :364/514A

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 364/514A; 382/10, 16, 19, 23, 28; 395/119, 120, 125, 127; 359/462, 466

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS (real-world scene; and model; and multiple cameras; and three-dimensional; and viewer criterion; and synthesizing; and spatial perspective)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US, A, 5,267,329 (ULICH ET AL) 30 November 1993, col. 9, lines 21-24, col. 7, lines 58-68, col. 8, lines 65-67, col. 8, lines 35-37, col. 8, lines 10-14, col. 10, lines 3-6.	1-36

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be part of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 JUNE 1996	16 SEP 1996
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer Emanuel Todd Voeltz Telephone No. (703) 305-9784

Form PCT/ISA/210 (second sheet)(July 1992)★